PB90-225731

ASA/EPA Conferences on Interpretation of Environmental Data: III. Sampling and Site Selection in Environmental Studies, May 14-15, 1987

(U.S.) Environmental Protection Agency, Washington, DC

1987

U.S. Department of Commerce
National Technical Information Service
NTIS

PB90-225731

United States
Environmental Protection
Agency

Office of Policy, Planning
and Evaluation
Washington, DC 20460

EPA-230-08-88-035

Statistical Policy Branch

EPA

ASA/EPA Conferences on Interpretation of Environmental Data

III Sampling and Site Selection In Environmental Studies May 14-15, 1987

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL
INFORMATION SERVICE
SPRINGFIELD, VA 22161

DISCLAIMER

This document has not undergone final review within EPA and should not be used to infer EPA approval of the views expressed.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank) 2. REPORT DATE 3. REPORT TYPE AND DATES COVERED

PB90 225731/AS 1987

4. TITLE AND SUBTITLE
ASA/EPA Conference on Interpretation of Environ. Data
III Sampling and Site Selection in Environ. Studies May 14-15, 1987

5. FUNDING NUMBERS

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Office of Policy, Planning and Evaluation
EPA
401 M St. SW, Wash. DC 20460

8. PERFORMING ORGANIZATION REPORT NUMBER
EPA-230-08-88-035

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Same as in item 7

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Release unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The eight papers and accompanying discussions in these proceedings are about drawing conclusions in environmental studies. The studies discussed involve air, water, gournd water, and soil. They employ techniques from various statistical areas such as probability sampling, reponse surface analysis, optimal design of experiments, time series analysis, spatial prediction, power transformations, and the analysis of variance. In the planning of an environmental study when almost all options are still open, most of these aspects are potentially relevant.

These proceedings are intended for statisticians involved in the planning of environmental studies. They should help the statistician anticipate the analysis to be performed.

No two authors have chosen the same aspect of environmental studies to examine. The diversity among authors suggests a major challenge. The challenge is to consider carefully each study aspect in the planning process.

14. SUBJECT TERMS

15. NUMBER OF PAGES
80

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNC	UNC	UNC	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	PR	- Project
G	- Grant	TA	- Task
PE	- Program Element	WU	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD	- See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE	- See authorities.
NASA	- See Handbook NHB 2200.2.
NTIS	- Leave blank.

Block 12b. Distribution Code.

DOD	- Leave blank.
DOE	- Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA	- Leave blank.
NTIS	- Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

PREFACE

This volume is a compendium of the papers and commentaries that were presented at the third of a series of conferences on interpretation of environmental data conducted by the American Statistical Association and the U.S. Environmental Protection Agency's Statistical Policy Branch of the Office of Standards and Regulations/Office of Policy, Planning, and Evaluation.

The purpose of these conferences is to provide a forum in which professionals from the academic, private, and public sectors exchange ideas on statistical problems that confront EPA in its charge to protect the public and the environment through regulation of toxic exposures. They provide a unique opportunity for Agency statisticians and scientists to interact with their counterparts in the private sector.

The conference itself and these proceedings are primarily the result of the efforts of the authors and discussants. The discussants not only provided their input to the proceedings but also reviewed the papers for the purpose of suggesting changes to the authors. The coordination of the conference and of the publication of the proceedings was carried out by Mary Esther Barnes and Lee L. Decker of the ASA staff. The ASA Committee on Statistics and the Environment was instrumental in developing this series of conferences.

The views presented in this conference are those of individual writers and should not be construed as reflecting the official position of any agency or organization.

Following the first conference, "Current Assessment of Combined Toxicant Effects," in May 1986, and the second conference, "Statistical Issues in Combining Environmental Studies," in October 1986, the third conference, "Sampling and Site Selection in Environmental Studies," was held in May 1987. One additional conference, "Compliance Sampling," was held in October 1987. A proceedings volume will also be published from this conference.

Walter Liggett, Editor
National Institute of Standards and Technology

INTRODUCTION

The eight papers and accompanying discussions in these proceedings are about drawing conclusions in environmental studies. These papers provide valuable guidance in the planning of future environmental studies.

The papers address many aspects of environmental studies. The studies discussed involve air, water, ground water, and soil. These studies are aimed at specific goals as diverse as the effect of a regulatory intervention, the design of a remediation effort, the spatial distribution of a hazardous material, the validity of an environmental model, and the impact of a power plant. Some studies emphasize in addition the planning of the field work and of the chemical analyses in the laboratory. The studies employ techniques from various statistical areas such as probability sampling, response surface analysis, optimal design of experiments, time series analysis, spatial prediction, power transformations, and the analysis of variance. In the planning of an environmental study when almost all options are still open, most of these aspects are potentially relevant.

These proceedings are intended for statisticians involved in the planning of environmental studies. Statistical planning is based on anticipation of the statistical analysis to be performed so that the necessary data can be collected. These proceedings should help the statistician anticipate the analysis to be performed. In addition, the papers discuss implications for planning new studies. No general prescriptions for planning are offered; none may be possible.

The emphases in these papers are quite different. No two authors have chosen the same aspect of environmental studies to examine. This diversity among authors who have all invested considerable time and effort in environmental studies suggests a major challenge. The challenge is to consider carefully each study aspect in the planning process. Meeting this challenge will require a high degree of professionalism from the statistician involved in an environmental sutdy.

Walter Liggett, Editor
National Institute of Standards and Technology

INDEX OF AUTHORS

Bailey, R. Clifton 73
Cressie, Noel A.C. 25
Englund, Evan J. 31
Folsom, Ralph E. 41
Hudak, G. 1
Jernigan, Robert W. 54
Johnson, W. Barnes 23
Liu, Lon-Mu 1
Livingston, Richard A. 55
Peterson, Bruce 70
Splitstone, Douglas E. 15
Stewart-Oaten, Allan 57
Thrall, Anthony D. 52
Tiao, George C. 1
Warren, John 40

TABLE OF CONTENTS

Preface iii

Introduction. WALTER S. LIGGETT, National Institute of Standards and Technology iv

Index of Authors v

I. THE STATISTICAL BASIS: RANDOMIZATION AND PROCESS CONTROL

A Statistical Assessment of the Effect of the Arizona Car Inspection/ Maintenance Program on Ambient CO Air Quality in Phoenix, Arizona. LON-MU LIU, G. HUDAK, GEORGE C. TIAO, University of Chicago 1

Sampling Design: Some Very Practical Considerations. DOUGLAS E. SPLITSTONE, IT Corporation 15

Discussion. W. BARNES JOHNSON, U. S. Environmental Protection Agency 23

II. INFERENCE ON CONTINUOUS SPATIAL DISTRIBUTIONS

Spatial Prediction and Site Selection. NOEL A. C. CRESSIE, Iowa State University 25

Spatial Autocorrelation: Implications for Sampling and Estimation. EVAN J. ENGLUND, U. S. Environmental Protection Agency 31

Discussion. JOHN WARREN, U. S. Environmental Protection Agency 40

III. DESIGNS BASED ON AUXILIARY INFORMATION

Sampling and Modeling Pollutant Plumes: Methods Combining Direct Measurements and Remote Sensing Data. RALPH E. FOLSOM, Research Triangle Institute 41

"Validation" of Air Pollution Dispersion Models. ANTHONY D. THRALL, Electric Power Research Institute 52

Modeling Pollutant Plumes. ROBERT W. JERNIGAN, The American University and Statistical Policy Branch, Environmental Protection Agency 54

Estimating the Spatial Uncertainty of Inferred Rates of Dry Acidic Deposition. RICHARD A. LIVINGSTON, University of Maryland 55

IV. STATISTICAL COMPARISON OF SITES

Assessing Effects on Fluctuating Populations: Tests and Diagnostics. ALLAN STEWART-OATEN, University of California-Santa Barbara 57

Comparisons with Background Environment: Strategies for Design. BRUCE PETERSON, CH2M Hill 70

Discussion. R. CLIFTON BAILEY, U. S. Environmental Protection Agency 73

Appendix A: Program 75

Appendix B: Conference Participants 77

A STATISTICAL ASSESSMENT OF THE EFFECT OF THE ARIZONA CAR INSPECTION/MAINTENANCE PROGRAM ON AMBIENT CO AIR QUALITY IN PHOENIX, ARIZONA

Lon-Mu Liu
G. Hudak
G. C. Tiao
University of Chicago, Graduate School of Business, 1101 E. 58th Street, Chicago, IL 60637

1. INTRODUCTION AND SUMMARY OF FINDINGS

This paper presents a statistical analysis of ambient carbon monoxide concentrations over the period 1971 to 1982 at three air monitoring sites in the state of Arizona. All three sites (Phoenix Central, Sunnyslope, and Phoenix South) are located in the Phoenix area where a vehicle inspection and maintenance (I/M or VEI) program has been in effect since January 1977. The principal objectives of this study are to assess the trend in the CO concentrations which can be associated with the Federal Motor Vehicle Emission Control Program to determine whether or not the I/M program in the Phoenix area has had a positive impact on the concentration level.

A summary of our principal findings is given in this section. Section 2 provides a description of the nature of the carbon monoxide, traffic and meteorological data used in the analysis. In Section 3 a preliminary trend analysis based on the CO readings alone is given. In Section 4 diurnal models relating CO to traffic and meteorological factors are constructed. Such models serve to identify the major exogenous variables affecting CO. In Section 5, time series intervention models for monthly means of CO readings are given. Traffic volume and relative humidity (a proxy for mixing height) are used as exogenous variables, and appropriate functional forms to model the effects of Federal emission standards and the I/M program are constructed. Finally in Section 6, a summary of the main results in trend analysis together with an assessment of the impact of the I/M program is presented.

Our principal findings are as follows:

(i) At all three sites one can observe a reduction in ambient CO concentration levels. The decrease at Phoenix Central is the largest. Considering the CO concentrations alone, this decrease is about 3.6% per year over the period 1973-1982. Further, the reduction at Phoenix Central is higher in the winter months, 5.7% per year, than in the summer months, 2.3% per year.

(ii) Based on models for the monthly means of CO at Phoenix Central which adjust for the effects of traffic changes and meteorological variations, the winter trend reduction is estimated at about 7.1% per year while the summer reduction is about 2.3% per year.

(iii) Yearly percentage changes in CO (based on adjusted monthly readings) at Phoenix Central are compared with two sets of emission factors derived from MOBILE3 analysis. One set of factors includes the expected effects of the I/M program and the other does not. It is found that, over the period 1973-1982, the year to year changes in CO concentrations in the winter months are in good agreement with changes in the set of emission factors which includes the effects of the I/M program. Provided that the emission factors are accurate, there is then some evidence from the observed ambient CO concentration levels to support the hypothesis that the I/M program has had a positive impact on ambient CO levels.

(iv) Analyses of diurnal models of CO over the period 1977-1982 produced trend estimates largely consistent with estimates based on monthly averages.

2. Data Base

2.1 The Arizona Inspection and Maintenance Program

The United States Clean Air Act Amendments of 1977 require that certain states implement vehicle inspection and maintenance programs (I/M programs) in certain of their major cities to reduce hydrocarbon (HC) and carbon monoxide (CO) emissions from gasoline powered vehicles. Arizona Revised Statutes §36-1772, which established the Vehicle Emissions Inspection (VEI) Program, requires that gasoline powered vehicles pass an annual inspection to ensure that their exhaust emissions meet standards established by the Department of Health Services (DHS). This program was initiated in 1976 on a trial basis in the Phoenix and Tucson areas, and repairs became mandatory in January 1977. The inspection requirment applies generally to gasoline powered vehicles which are less than 14 years old and located within designated portions of Pima and Maricopa Counties which do not meet the carbon monoxide standards of the Federal Clean Air Act. According to the VEI program vehicles are tested annually to ensure that carbon monoxide and hydrocarbons in their exhaust emissions meet standards established by DHS. Motorists whose vehicles fail to meet these standards must repair their vehicles and submit to a retest.

Data on the annual number of inspections and failure rates obtained from the Arizona DHS. Generally speaking, the monthly failure rates between 1977 and 1983 range between 15 and 25 percent, and they are higher in 1979 and 1980 than in other years.

2.2 Carbon Monoxide Data (CO)

The data consists of hourly CO concentrations (ppm) recorded at three Phoenix sites. The Phoenix CO measurment locations are: Phoenix Central (at 1845 E. Roosevelt in downtown), Sunnyslope (at 8531 N. 6th Street), and Phoenix South (at 4732 S. Central).

The hourly data on CO concentrations at these three stations vary in length: Phoenix Central (71/1 - 82/12), Sunnyslope (74/10 - 82/12), and Phoenix South (74/10 - 82/12). Missing data occur at all three sites. Phoenix Central has one month (January, 1979) in which data are completely missing, the other two sites have several months completely missing.

2.3 Traffic (TR)

Many studies have shown that ambient CO concentrations are approximately proportional to traffic (Tiao, Box and Hamming (1975), Tiao and Hillmer (1978), Ledolter and Tiao (1979b), Tiao, Ledolter and Hudak (1982)). It is, therefore, necessary to incorporate changes in traffic into the trend analysis. Ideally, one would want to use traffic data recorded throughout the area affecting CO measurement sites over the entire period under study. Unfortunately, such detailed data are not available for this study.

Estimates of relative traffic volume per day in the vincinity of Phoenix and Central and Phoenix South between 1970 and 1983 were provided by the Bureau of Air Quality Control (BAQC), Arizona Department of Health Services. These figures are listed in Table 2.1 Adjustment fractions for different months of a year, different days of a week, and different hours of a day were also provided by the Bureau of Air Quality Control and are listed in Table 2.2.

2.4 Meteorological Data

Apart from traffic, variations in ambient CO concentrations are to a large extent affected by meteorological conditions:

(a) wind speed and wind direction affect the transport and difussion of CO emissions (with low wind speed resulting in high CO levels);

(b) inversion (mixing heights) affects the volume of air available for dilution of CO emissions;

(c) temperature, solar intensity and relative humidity are related to the duration and intensity of temperature inversions and the degree of vertical mixing;

(d) meteorological variables (such as temperature) influence the efficiency factors of car engines (cold starts leading to higher CO levels). Thus, one should consider incorporating these variables as exogenous factors in a trend analysis of CO.

Meteorological data were obtained from the National Climatic Center (NCC) for the monitoring station of the National Weather Service in Phoenix (near Phoenix Central). Data on wind speed (WS, knots), wind direction (WD, tens of degree), temperature (TP, °F), relative humidity (RH, percent) and precipitation frequency (PREC, 0-60) were obtained for the time period 1971/1 - 1982/12. Unfortunately, mixing height data were not available. A closely related variable whose data was available is delta temperature (ΔT). The ΔT variable is the difference in temperature readings recorded at different heights. Since ΔT is a measure of atmospheric stability, it may be used as a proxy for mixing heights. Hourly ΔT measurements were made by the BAQC at the 6th Street and Bultler, where

ΔT = temperature (^{0}F) at 30 feet
- temperature (^{0}F) at 8 feet

The hourly ΔT data are available between 77/5 - 82/12 but with many missing days and months.

3. Preliminary Trend Analysis of CO

The primary objective of this study is to assess statistical evidence of the impact of the Arizona I/M program on the ambient CO concentration levels. Our approach to this problem is as follows. We first present in this section a preliminary trend analysis based on the CO data alone. The effects of traffic changes and the influences of meteorological variables are then considered in the next two sections (Sections 4 and 5) where various models relating CO at Phoenix Central to these exogenous factors are constructed. Finally, in Section 6, we compare the model-based trend estimates with the expected reduction in CO emissions based on EPA MOBILE3 analysis applied to the Phoenix Central location with and without the I/M program.

3.1 Preliminary CO Trend Analysis for Phoenix Central

Figure 3.1 shows monthly means of CO at Phoenix Central. It is seen that there was an apparent down trend in CO over the period 1973-1982. (We have been informed by the DHS that CO data for the period 1971/1 - 1972/3 received by us are incorrect and hence they have not been used in our analysis.) As a first approximation, if we assumed a linear time trend operating from 1973-1982, then the estimated CO reduction in the yearly means would be about 3.6% per year. Further study of the data shows that the percentage reduction is higher in the winter months (Oct. - Feb.) than in the summer months (April - August). Specifically, over the period 1973-1982, the decrease is about 2.3% per year in the summer months and 5.7% per year in the winter months.

3.2 Preliminary CO Trend Analysis for Sunnyslope and Phoenix South

For the Sunnyslope location, Figure 3.2 shows that over the period 1974-1982 there are many months for which data were missing. If we exclude the four years containing months of missing data (i.e., 1974, 1979, 1980 and 1982), we see that the reduction in CO is smaller than that at the Phoenix Central location. In particular, the estimated reduction in the yearly means of CO here is about 2.6% per year.

Similarly, for the Phoenix South site, we see from Figure 3.3 that there are gaps of missing data. Excluding the three years 1974, 1979 and 1980, the estimated reduction in the yearly means of CO is about 2.4% per year, which is again smaller than that at the Phoenix Central site.

3.3 General Remarks

The preliminary trend calculations were based on the CO readings alone. Recall from Table 2.1 that the traffic volume increased steadily over the years 1970-1982 in the Phoenix area. Thus, the estimated reduction in CO emissions would be higher than the figures given above when the traffic increases are factored into the analysis. Also, the preceding analysis did not take into account the influence of meteorological variables. We now turn to discuss various moels relating CO

to exogenous traffic and meteorological factors. In view of the fact that the Phoenix Central site has the longest and most complete data record on CO and that the concentration level is also much higher there than those of the two other sites, we shall confine ourselves to the Phoenix Central data in our modeling study.

4. Diurnal Models of CO

The focus of this section is on modeling the diurnal behavior of CO which serves to identify the main factors affecting CO, and thus motivates the trend models discussed in Sections 5 and 6.

As mentioned in Section 2, detailed traffic information, such as hourly traffic counts, was not available for the area influencing the CO measurement sites. In what follows, we use the estimated relative traffic volume over 1971-1982 pertaining to Phoenix Central and the adjustment fractions for different months, day and hours of a day provided by BAQC (see Tables 2.1 and 2.2).

We have found that analyses using logarithmic transformation of the variables seem to be more meaningful than those using the original variables. The following notations will be employed:

Carbon Monoxide	$LCO = \ln(CO + 0.25)$
Traffic	$LTR = \ln(TR)$
Inverse of Wind Speed	$LIW = \ln(1/(WS + 0.25))$
Temperature	$LTP = \ln(TP)$
Relative Humidity	$LRH = \ln(RH)$
Delta Temperature	$L\Delta T = \ln(\Delta T + c)$, $c = 2.0$ for winter seasons, $c = 4.0$ for summer seasons

4.1 Formulation of the Models

In general, we can write a model relating CO to the input traffic and meteorological variables as

$$CO = g(TR)f(met.)$$

where g(TR) is a function of the input traffic and f(met.) is a function of meteorological variables. The function g(TR) will depend, among other things, on the units of measurements of CO and TR. As an approximation it seems reasonable to suppose that

$$g(TR) = k\, TR^{\beta_1} \qquad (4.1)$$

where k is a constant measuring the CO emissions and β_1 measures the percentage change in CO due to a one percent change in TR.

Upon studying the diurnal diagrams of CO vs various meteorological variables, we have found that the major meteorological factors influencing the diurnal behavior of CO are ΔT and the wind speed (WS) or its inverse $IW = WS^{-1}$. We may also approximate the function f(met.) as

$$f(met.) = c(\Delta T)^{\beta_2}\, (IW)^{\beta_3}$$

so that we have a model of the multiplicative form

$$CO = \omega\, TR^{\beta_1} (\Delta T)^{\beta_2} (IW)^{\beta_3} \qquad (4.2)$$

where $\omega = ck$. For trend assessment, we can make the constant k to be dependent on time, and in particular, we may write

$$k = k_0\, e^{-k_1 T} \qquad (4.3)$$

where k_0 is a constant, T measures the time unit (year) under consideration and k_1 the percentage change in CO emissions per unit time.

We are thus led to a model of the multiplicative form

$$CO = \omega\, TR^{\beta_1} (\Delta T)^{\beta_2} (IW)^{\beta_3} u$$

where u is the error term, and upon taking logarithms,

$$LCO = \beta_0 + \beta_1 LTR + \beta_2 L\Delta T + \beta_3 LIW + \varepsilon \qquad (4.4)$$

where $\beta_0 = \ln(\omega)$ and $\varepsilon = \ln(u)$. For diurnal data, let t stands for the t-th hour of the day, t=1,...,24. Because of the dynamic nature of the traffic and meteorological effects, we would expect that LCO_t could be related to some linear aggregates of the past values of these exogenous variables and also that the error term ε_t would be serially correlated. As an approximation, we consider the diurnal model

$$LCO_t = \beta_0 + \beta_1 \overline{LTR}_{t-1}^{(\phi)} + \beta_2 \overline{L\Delta T}_{t-1}^{(\phi)} + \beta_3 \overline{LIW}_{t-1}^{(\phi)} + \varepsilon_t \qquad (4.5)$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + a_t$$

where a_t's are white noise with zero means and common variance σ^2.

$$\overline{LTR}_{t-1}^{(\phi)} = (1-\phi)[LTR_{t-1} + \phi LTR_{t-2} + \phi^2 LTR_{t-3} + \ldots]$$

and similarly for $\overline{L\Delta T}_{t-1}^{(\phi)}$ and $\overline{LIW}_{t-1}^{(\phi)}$.

For parameters estimation, we can write (4.5) in the alternative form

$$LCO_t = \beta_{0*} + \phi LCO_{t-1} + \beta_{1*} LTR_{t-1} + \beta_{2*} L\Delta T_{t-1} + \beta_{3*} LIW_{t-1} + a_t \qquad (4.6)$$

where $\beta_{j*} = \beta_j(1-\phi)$, $j = 0,\ldots,3$. This is in the form of a linear regression model and hence the parameters β_{j*}'s and ϕ can be estimated from standard least squares routines.

To distinguish the behavior of CO between the summer and the winter, we have estimated the model (4.6) separately for each of these two seasons employing all available data over the years. For each season, we let $LCO^{(i)}$, $LTR^{(i)}$, $L\Delta T^{(i)}$, $LIW^{(i)}$ represent the average readings of these variables at the t-th hour for the i-the year, and let $\beta_{0*}^{(i)}$ be a separate constant for each of the years considered. Thus

$$LCO_t^{(i)} = \beta_{0*}^{(i)} + \phi LCO_{t-1}^{(i)} + \beta_{1*} LTR_{t-1}^{(i)} + \beta_{2*} L\Delta T_{t-1}^{(i)} + \beta_{3*} LIW_{t-1}^{(i)} + a_t^{(i)} \qquad (4.7)$$

$$i=1,2,\ldots, \qquad t=1,\ldots,24$$

For the winter diurnal model, only data in 1977/78, 1979/80, 1980/81 and 1981/82 are complete. The traffic $TR^{(i)}$ is the diurnal traffic pattern listed in Table 2.2 divided by the annual traffic volume in that year, and then multiplied by the annual traffic volume in 1972 for normalization. For the summer diurnal model, data from 1977 to 1982 are complete. Note that in model (4.7), the term $\beta_{0*}^{(i)}$ may vary from year to year, but the coefficient ϕ, β_{1*}, β_{2*}, and β_{3*} are held constant over the years.

For the convenience of trend assessment, we may impose a simple linear time function on the yearly constants $\beta_{0*}^{(i)}$, such as

$$\beta_{0*}^{(i)} = \alpha_0 + k_{1*} i \qquad (4.8)$$

Thus the model in (4.7) can be written as

$$LCO_t^{(i)} = \alpha_0 + k_{1*} i + \phi LCO_{t-1}^{(i)} + \beta_{1*} LTR_{t-1}^{(i)} + \beta_{2*} L\Delta T_{t-1}^{(i)} + \beta_{3*} LIW_{t-1}^{(i)} + a_t^{(i)} \qquad (4.9)$$

Note that from (4.3), here we have i=T, $k_{1*} = k_1(1-\phi)$ and the coefficient k_1 is the percentage reduction of CO from year to year.

4.2 Fitting Results

The models in (4.7) and (4.9) are fitted by least squares to hourly averages of winter and summer separately. Figures 4.1(a) and (b) plot the predicted values of $LCO^{(i)}$ from the fitted models vs the actual observations for the winter and the summer seasons over the years considered, showing that in both cases the models give an excellent representation of the CO data, especially in the winter season.

The parameter estimates are summarized in Table 4.1 where the associated estimated standards errors are given in the parentheses. It is clear that CO concentration is positively associated with TR, ΔT and IWS as to be expected, but in the summer the effect of wind speed is very slight. For the winter season, the estimated constants $\beta_{0*}^{(i)}$ decreased steadily from 2.260 in 1977/78 to 1.867 in 1981/82. The slope coefficient k_{1*} is estimated to be -.088 so that in terms of

$k_1 = k_{1*}/(1-\phi) = -.088/(1-.408) = -.149$

with an estimated standard error of .042. This is saying that in the winter seasons, the average reduction in CO over the year 1977/78 to 1981/82 is approximately 15% (with an estimated standard error of 4%) per year. On the other hand, for the summer season, the constants $\beta_{0*}^{(i)}$ vary a great deal over the years and no simple linear down trend can be established from the data. Note that, of course, CO concentrations are substantially higher in the winter seasons.

5. Time Series Models for Monthly Averages of CO at Phoenix Central

In this section, we consider time series models relating monthly means of CO at the Phoenix Central site to traffic and meteorological variables. The objective is to assess the trends in the CO data. In what follows we shall employ data from October 1972 onward.

5.1 Formulation of the Trend Models

The reduction in the CO concentrations which was reported in the preliminary trend analysis of Section 3 as well as in the diurnal model analysis of Section 4 can be attributed to several factors:

(1) Federal new car CO Emission Standards:

Over the last 15 years the federal CO emission standards became increasingly more stringent. For example, the CO emissions were limited to 34g/travelled mile for 1970/71 model year cars to 28g/mile for 1972-74 model year cars, and to 15g/mile for 1975-77 model year cars. Since the CO concentrations in the winter and summer are quite different, one may suspect the trend reduction will also be different. In the models to be discussed, we have incorporated such possibility.

(2) The Phoenix I/M Program: The Phoenix vehicle inspection and maintenance program is expected to have an additional impact on the CO emissions. We may examine this impact using intervention analysis, Box and Tiao (1975). One would certainly expect that the effect due to the I/M program will be confounded with the trend reduction due to Federal emission standards. To access the impact of the I/M program more objectively, in this section the overall trends in the winter and summer are first estimated using time series models. In Section 6 we will then compare these trends with the expected reduction in emissions resulted from MOBILE3 analysis.

(3) Traffic: It was discussed in Section 4 that CO concentrations can be expected to be proportional to traffic. To adjust for the increase in traffic volumes over the years considered, the monthly CO concentrations are divided by the annual traffic volume in that year (see Table 2.1) and then multiplied by 208.275 for normalization. This adjusted CO is denoted as CO_t^*, where

$$CO_t^* = (CO_t/TR^{(i)}) * 208.275 , \qquad (5.1)$$

$$LCO_t^* = ln(CO_t^*) , \qquad (5.2)$$

and t stands for a month.

For the purpose of comparison, we also analyze the monthly CO concentration without adjusting for the traffic changes.

(4) Meteorological Variables: A functional form relating CO concentrations to meteorological variables was considered in Section 4. The model was of the form

$$f(met.) = c(\Delta T)^{\beta_2} (IW)^{\beta_3} . \qquad (5.3)$$

Since data on ΔT were not available in many months, we examined models which incorporate relative humidity (RH), wind speed (WS) and temperature (TP). These variables are used as possible proxies for ΔT. Of these models we found that a simple model of the form

$$LCO_t^* = \beta_0 + \gamma LRH_t \qquad (5.4)$$

led to a good description of the data.

(5) Seasonality: We saw earlier in Figure 3.1 that CO concentrations exhibit a strong seasonal pattern, with high concentrations in the winter and low values in the summer. This seasonal variation is largely explained by changes in meteorological variables. From (5.4), we may write

$$LCO_t^* = \beta_0 + h(t) + \gamma LRH_t + \eta_t .$$

where h(t) is a function of time t representing trend movement in the data and η_t is the noise term. It is found that the noise term can be simply expressed as

$$\eta_t = \omega \sin(2\pi t/12) + a_t .$$

where $\omega\sin(2\pi t/12)$ is a deterministic sine wave of period 12 representing the seasonal behavior of CO not explained by that in RH. Combining the above factors, the following two time series models for the monthly means are considered:

$$Y_t = \beta_{01} + \beta_{02} WN_t + \sum_{i=1}^{9} \alpha_{1i} IDS_t^{(i)} + \sum_{j=1}^{9} \alpha_{2j} IDW_t^{(j)} + \gamma LRH_t + \omega \sin\left(\frac{2\pi t}{12}\right) + a_t \quad (5.5)$$

or

$$Y_t = \beta_{01} + \beta_{02} WN_t + \beta_1 TS_t + \beta_2 TW_t + \gamma LRH_t + \omega \sin\left(\frac{2\pi t}{12}\right) + a_t \quad t=1,2,\ldots \quad (5.6)$$

where Y_t: LCO_t, logarithm of $(CO_t + 0.25)$, or LCO_t^*, logarithm of $((CO_t + 0.025)/TR^{(i)})$ *208.275, i = 1972,...,1982

WN_t: winter months (October-February) indicator, i.e.,
WN_t = 1 if it is a winter month
= 0 otherwise

$IDS_t^{(i)}$: summer months (April-August) indicator for the i-th year, i.e.,
$IDS_t^{(i)}$ = 1 if it is a summer month in the i-th year
= 0 otherwise
i=1 for 1974,..., i=9 for 1982

$IDW_t^{(j)}$: winter month indicator for the j-th year, i.e.,
$IDW_t^{(i)}$ = 1 if it is a winter month in the j-th year
= 0 otherwise
j = 1 for 1973/74,..., j=9 for 1981/82

TS_t: summer trend
TS_t = 1 for all summer months in 1973
TS_t = 2 for all summer months in 1974
⋮
TS_t = 10 for all summer months in 1982
TS_t = 0 otherwise

TW_t: winter trend
TW_t = 1 for all winter months in 1972/1973
TW_t = 2 for all winter months in 1973/1974
⋮
TW_t = 10 for all winter month in 1981/1982
TW_t = 0 otherwise

In the above two models, Model (5.5) provides more detailed information of the change of CO concentrations from year to year. Specifically, α_{1i} measures the effect (percentage change) of CO in the i-th summer compared with that for the base year 1973, and α_{2j} the effect (percentage change) of CO in the j-th winter compared with that of the 1972/73 winter. Instead, Model (5.6) assumes linear time trends and β_1 and β_2 represent the percentage reductions of CO concentrations per year in the summer and winter respectively.

5.2 Model Estimation

The parameters in models (5.5) and (5.6) are fitted to monthly averages of CO, CO^* and RH using the SCA statistical system (Liu et al, 1983). The estimation results for Phoenix Central are listed in Table 5.1 for LCO_t and Table 5.2 for LCO_t^*.

5.3 Trend Analysis for Phoenix Central

For LCO_t (without normalization by the traffic volumes), Figuere 5.1(a) gives a plot of the estimated effects α_{2j} (compared with the based year, winter of 1972/73) for the winters of 1973/74 to 1981/82, and Figure 5.1(b) presents a plot of α_{1i} (compared with the base year, summer of 1973) for the summers of 1974 to 1982. Similarly, Figures 5.2(a) and (b) give the corresponding plots of LCO_t^* (after traffic normalization). From these figures and the estimates in Tables 5.1 and 5.2, we make the following observations:

(i) Both in the winters and in the summers, the down trend in CO is more pronounced after traffic normalization. This is, of course, to be expected since the traffic volume was steadily increasing throughout the period 1973-1982 considered.

(ii) By comparing Figures 5.1(a) with 5.2(a), and Figures 5.1(b) with 5.2(b) we see that the relative changes in the yearly effects are in fact much closer between LCO_t and LCO_t^* from 1976 onward compared with the changes before 1976. This is also reasonable since the annual percentage increase in traffic volume was much larger before 1976 - see Table 2.1. Thus, normalization by traffic volume has its major impact on the trend movement during the early part of the data period considered.

(iii) For the normalized LCO_t^* data, model (5.6) yields an estimated down trend of 7.1% per year in the winter season over the entire period 1972/73 to 1981/82. The magnitude of the percentage reduction in the summer season is much smaller, 2.3% per year. Also, figures 5.2(a) and (b) show that the summer effects exhibit considerably more variability than the winter effects do from year to year. In Section 6, we shall compare these estimated effects with the expected reduction in emissions from EPA MOBILE3 analysis.

6. Summary of Trend Analysis and Assessment of the Impact of the I/M Program at Phoenix Central

In Sections 3-5, we have performed several types of trend analyses on the CO concentration levels. Our main focus has been on CO at Phoenix Central where the readings are substantially higher than those at the other two sites, Sunnyslope and Phoenix South, and also where the data span is the longest and most complete of these three locations. The main results at Phoenix Central can be partially summarized as follows:

(i) Over the time period April 1972 to December 1981 studied, there is clear evidence of a down trend in the CO concentrations

(ii) The down trend is more pronounced and smoother in the winter months (October-February), while considerably higher variability exists in the trend movement in

the summer months (April-August). Note that CO concentration levels are substantially higher in the winter.

(iii) For the winter months over the period 1972/1973 to 1981/1982:

(a) preliminary analysis of monthly averages of CO readings alone yields an estimated trend reduction of 5.6% per year

(b) after allowing for traffic and meteorological changes, our analysis of the monthly CO averages shows an reduction of about 7.1% per year.

(iv) Over the four winters 1977/78, 1979/80, 1980/81 and 1981/82, our diurnal model relating CO to traffic and meteorological factors yield an estimated reduction of 15% per year.

6.1 Assessment of the Impact of the I/M Program

Empirical assessment of the impact of the Arizona I/M program on CO concentration levels is made difficult by (i) the confounding between the expected effect of Federal emission standards and that of the I/M program and (ii) the lack of a comparable "control" site without the I/M program. One approach to this problem would be to construct a time series model of the monthly averages of CO with a linear time function characterizing the general down trend in the data and an additional term representing possible change in level of slope at the inception of the I/M program. For examples, we may contemplate the models

$$CO_t = \alpha_0 + \alpha_1 t + \alpha_2 I_t^{(t_0)} + N_t \qquad (6.1)$$

or

$$CO_t = \alpha_0 + \alpha_1 t + \alpha_2 (t-t_0) I_t^{(t_0)} + N_t \qquad (6.2)$$

where t_0 signifies the inception of the I/M program,

$$I_t^{(t_0)} = \begin{cases} 0 & t < t_0 \\ 1 & t \geq t_0 \end{cases}$$

and N_t us the noise term which could be auto-correlatated. In (6.1) α_2 measures the level change and in (6.2) α_2 measures the slope change associated with the I/M program. One may then estimate the α's from the data and test the significance of α_2. This approach has been adopted by McCleary and Nienstedt (1983) in their study of the Arizona CO data, and they have concluded that the I/M program has had no impact because the estimated α_2 from models of the form (6.1) or (6.2) is not statistically significant. Their results can be largely seen from the plots of the estimated yearly effects of LCO_t^* in Figure 5.2. Consider, for example, the winter situation. Clearly there is little evidence from these yearly effects to support either a change of level or slope beginning in 1977/78.

In the absence of any information about the magnitude of expected effects of either the Federal Emission standard and/or the I/M program, and assuming that the linear time functions in (6.1) and (6.2) are suitable approximations to the shapes of the expected impacts of these two types of control measures, the above approach would be useful. On the other hand, when accurate information about the magnitude of the expected effects is available, other formulations of the problem are possible.

Table 6.1 lists the CO emission factors from 1972 to 1982 with and without the I/M program for the central area of Phoenix supplied to us by the DHS. These factors were derived from EPA MOBILE3 analysis with adjustment for vehicle tampering/misfueling and varying speed over years. The yearly factors correspond to estimates for January 1 of the years. Assuming that these factors accurately represent the expected effects of the Federal emission standards and the expected impact of the I/M program, we can then employ the CO readings to assess to what extent trends in the emission factors with and without I/M are compatible with evidence from the data. For this analysis, we shall use monthly averages of the CO readings as discussed in Section 5.

Figure 6.1(a) shows plots of the logarithms of the emission factors over time where a dot " " corresponds to the situation without I/M and a cross "x" to the situation with I/M effects included. The points in this figure then represent the percentage changes in emissions from year to year with and without the effects of the I/M program. Superimposed in the same figure by the circles "0" are the yearly effects of LCO_t^* for the winter months estimated from model (5.5) and shown earlier in Figure 5.2(a). It is clear that the trends in the estimated yearly effects are in reasonable agreement with emission factors with the I/M program.

In a similar way, Figure 6.1(b) shows logarithms of the emission factors together with the yearly effects of LCO_t^* for the summer seasons shown earlier in Figure 5.2(b). In this case, the estimated yearly effects are in better agreement with emission factors without the I/M program. However, it should again be noted that CO readings in the winter months are substantially higher and that the fluctuations in the summer yearly effects are much larger.

To formalize the analysis, we let $f_1(t)$ be a function whose values are given by the logarithms of the emission factors with the I/M program (i.e., dots from 1972 to 1977 and crosses afterwards) and $f_0(t)$ be a function corresponding to the situation without the I/M (i.e., dots from 1972 to 1982). We now replace the yearly effects components in (5.5) by $f_1(t)$ and $f_0(t)$ yielding the following two models

$$LCO_t^* = \beta_{01} + \beta_{02} WN_t + \beta_{S1} f_1(t)\delta_S(t) + \beta_{W1} f_1(t)\delta_W(t) + \gamma LRH_t + \omega \sin\left(\frac{2\pi t}{12}\right) + a_t \qquad (6.3)$$

$$LCO_t^* = \beta_{01} + \beta_{02} WN_t + \beta_{S0} f_0(t)\delta_S(t) + \beta_{W0} f_0(t)\delta_W(t) + \gamma LRH_t + \omega \sin\left(\frac{2\pi t}{12}\right) + a_t \qquad (6.4)$$

where $\delta_S(t) = 1$ t in summer months, 0 otherwise

$\delta_W(t) = 1$ t in winter months, 0 otherwise

Thus, the model (6.3) aims at assessing to what extent $\theta_{S1} = 1$ and $\theta_{W1} = 1$ are supported by the data (i.e., the "with I/M" hypothesis), whilst (6.4) is intended for

$$\theta_{SO} = 1 \text{ and } \theta_{WO} = 1$$

(the "without I/M" hypothesis). Estimation of (6.3) and (6.4) gives

		estimate	s.e.
Model (6.3)	θ_{S1}	.26	(.18)
	θ_{W1}	.97	(.21)
	σ^2error	.071	
Model (6.4)	θ_{SO}	.83	(.48)
	θ_{WO}	2.41	(.56)
	σ^2error	0.73	

The evidence is clearly in favor of $\theta_{W1} = 1$ (supporing the emission factors with I/M effects for the winter months), and, to a much lesser extent, is in favor of $\theta_{SO} = 1$ (without I/M effects during the summer months).

One may also proceed one step further by combining the two emission factor components in (6.3) and (6.4) into a single overall model and perform various hypothesis tests concerning

$$\theta_{W1},\ \theta_{WO},\ \theta_{S1}, \text{ and } \theta_{SO},$$

but the general conclusions are similar.

6.2 Sensitivity of Trend Analysis to Traffic Normalization

So far in Sections 5-6, we have accounted for the effect of traffic changes by normalizing the CO readings using the relative volumes in Tables 2.1 and 2.2. This has been done by assuming that CO is proportional to the traffic volume. While this seems reasonable, our diurnal model (4.9) suggests that, for the winter months, CO is proportional to $(TR)^{\beta_1}$ where β_1 is estimated to be .266/(1-.408)=.45. Thus, one should perhaps normalize the CO by $(TR)^{1/2}$, the square root of the traffic volume. We have conducted trend analyses using

$$LCO_t^* = \ln((CO_t+0.25)/TR_t^{1/2}) \cdot 208.275$$

similar to those given in Sections 5-6, little changes in the main conclusions. In particular, for the models in (6.3) and (6.4), we have obtained the following results

		estimate	s.e.
Model (6.3)	θ_{SO}	.55	(.46)
	θ_{WO}	2.07	(.55)
	σ^2error	.068	
Model (6.4)	θ_{SO}	.55	(.46)
	θ_{WO}	2.07	(.55)
	σ^2error	.069	

Thus the evidence is still supportive of the emission factors with I/M effects for the winter months.

Finally, we note that an extensive exploratory data analysis of the CO, traffic and meteorological variables, trend assessments of morning and evening peak hour CO readings, and further details of the analyses reported in this paper can be found in Tiao, Liu and Hudak (1984).

References

Box, G.E.P. and Tiao, G.C. (1975), "Intervention Analysis with Application to Economic and Environmental Problems", Journal of the American Statistical Association 70, 70-79.

Ledolter, J. and Tiao, G.C. (1979a), "A Statistical Analysis of New Jersey CO Data", in Proceedings of APCA ASQC Conference on Quality Assurance in Air Pollution Management, 282-293.

Ledolter, J. and Tiao, G.C. (1979b), "Statistical Models for Ambient Air Pollutants, with Special Reference to the Los Angeles Catalyst Study (LACS) Data", Environmental Science and Technology 13, 1233.

Liu, L.M., Hudak, G.C., Box, G.E.P., Muller, M.E. and Tiao, G.C. (1983), The SCA System for Univariate-Multivariate Time Series and General Statistical Analysis, Scientific Computing Associates, P.O. Box 625, DeKalb, IL 60115.

McCleary, R. and Nienstedt B.C. (1983), "Time Series Analysis of the Impact of the Vehicle Emissions Inspection Program (I/M) on Ambient Air Quality in Phoenix and Tucson", Report to the Auditor General, State of Arizona.

Tiao, G.C., Box, G.E.P. and Hamming, W.J. (1975), "A Statistical Analysis of the Los Angeles Ambient Carbon Monoxide Data 1955-1972", Journal of Air Pollution Control Association 25, 1130-1136.

Tiao, G.C. and Hillmer, S.C. (1978), "Statistical Models for Ambient Concentrations of Carbon Monoxide, Lead and Sulfate Based on the LACS Data", Environmental Science and Technology 12, 820.

Tiao, G.C., Ledolter, J. and Hudak, G.C. (1982), "Statistical Analysis of the Effect of Car Inspection and Maintenance Programs on the Ambient CO Concentrations in Oregon", Environmental Science and Technology 16, 328.

Tiao, G.C., Liu, L.M., and Hudak, G.B. (1984). "Statistical Analysis of Aerometric Data to Assess the Effect of the Arizona I/M Program on Ambient Quality in Phoenix, Arizona", Final Report to EPA under Contract #68-02-3513.

Acknowlegement: This research was supported by EPA Contract #68-02-3513

Table 2.1
Relative Traffic Volume Per Day in Phoenix Area

Year	Phoenix Central	Phoenix South
70	134.000	17.733
71	151.500	18.467
72	168.400	19.200
73*	181.900	19.933
74	193.00	20.667
75*	202.200	21.400
76	208.300	22.133
77*	212.000	22.867
78	214.800	23.600
79*	216.400	24.333
80	217.900	25.066
81	218.700	25.799
82*	219.300	26.533
83	219.700	27.266

(*): interpolated data

Table 2.2
Adjustment Fractions
for Months, Days and Hours

Monthly	Adjustment	Month	Adjustment
Jan.	1.02	July	.94
Feb.	1.04	Aug.	.96
Mar.	1.03	Sep.	.98
Apr.	1.02	Oct.	1.00
May	.97	Nov.	1.02
June	.96	Dec.	1.05

Day of Week	Adjustment
Mon.	.96
Tues.	.96
Weds.	.98
Thurs.	1.00
Fri.	1.09

Hour of Weekday Adjustments

Hr.	Adjustment	Hr.	Adjustment
1	.011	13	.060
2	.007	14	.059
3	.004	15	.064
4	.004	16	.072
5	.005	17	.081
6	.012	18	.076
7	.035	19	.056
8	.069	20	.045
9	.057	21	.037
10	.051	22	.034
11	.054	23	.028
12	.059	24	.020

Figure 3.1

Plot of Monthly CO Means at Phoenix Central

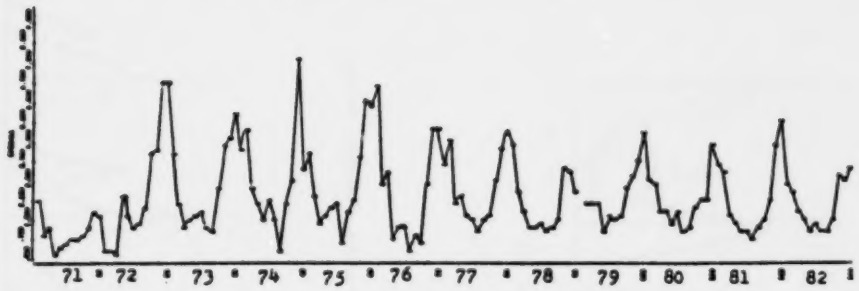


Figure 3.2

Plot of Monthly CO Means at Sunnyslope

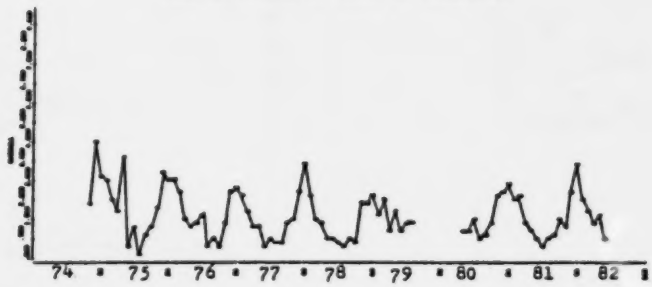


Figure 3.3

Plot of Monthly CO Means at Phoenix South

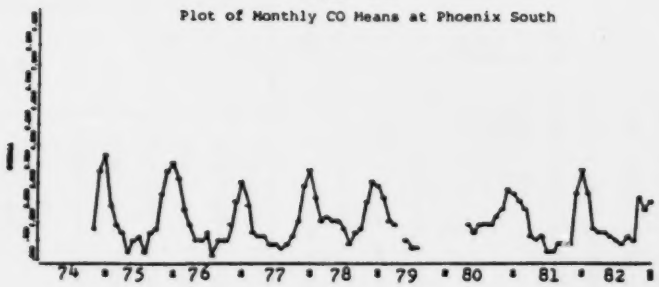


Table 4.1

Estimation Results for the Diurnal Models (4.7) and (4.9)

(a) Winter Seasons (1977/78 - 1981/82)

Model (4.7)		Model (4.9)	
$\beta_{0a}^{(1)}$ (1977/78)	2.260 (.369)		
$\beta_{0a}^{(3)}$ (1979/80)	2.161 (.344)		
$\beta_{0a}^{(4)}$ (1980/81)	1.958 (.318)	α_a	2.163 (.357)
$\beta_{0a}^{(5)}$ (1981/82)	1.867 (.275)	k_{1a}	-.088 (.025)
ϕ	.408 (.058)	ϕ	.435 (.057)
β_{1a}	.266 (.028)	β_{1a}	.258 (.028)
β_{2a}	.407 (.040)	β_{2a}	.401 (.040)
β_{3a}	.641 (.157)	β_{3a}	.553 (.138)
$\hat{\sigma}$	.194	$\hat{\sigma}$	.196

(b) Summer seasons (1977 - 1982)

Model (4.7)		Model (4.9)	
$\beta_{0a}^{(1)}$ (1977)	.284 (.389)		
$\beta_{0a}^{(2)}$ (1978)	.123 (.374)		
$\beta_{0a}^{(3)}$ (1979)	.292 (.373)		
$\beta_{0a}^{(4)}$ (1980)	.197 (.343)		
$\beta_{0a}^{(5)}$ (1981)	.138 (.297)	α_a	.030 (.375)
$\beta_{0a}^{(6)}$ (1982)	.183 (.288)	k_{1a}	-.002 (.024)
ϕ	.467 (.069)	ϕ	.530 (.064)
β_{1a}	.117 (.035)	β_{1a}	.102 (.034)
β_{2a}	.386 (.051)	β_{2a}	.357 (.050)
β_{3a}	.086 (.148)	β_{3a}	.001 (.131)
$\hat{\sigma}$	.250	$\hat{\sigma}$	.252

Figure 4.1(a)

Observed (A) vs predicted (B) hourly CO averages of Phoenix Central (winter season)

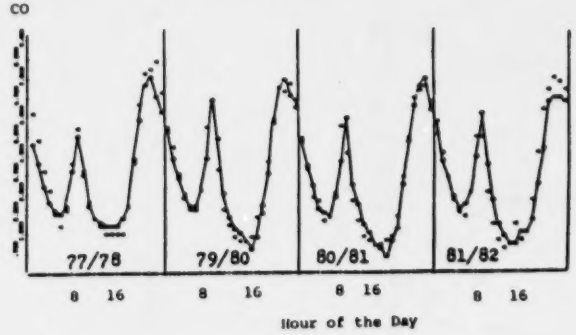


Figure 4.1(b)

Observed (A) vs predicted (B) hourly CO averages of Phoenix Central (summer season)

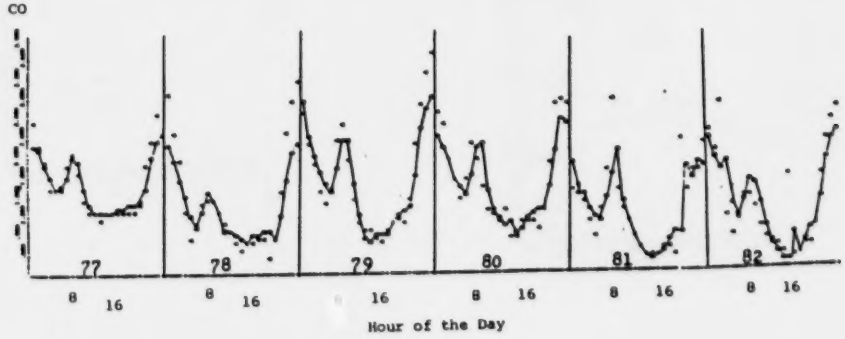


Table 5.1

Estimation Results for the Time Series Models (5.5) and (5.6)

$Y_t = LCO_t$ (10/72 - 9/82)

	Model (5.5)			Model (5.6)	
	β_{01}	2.997 (.367)		β_{01}	2.850 (.358)
	β_{02}	.447 (.159)		β_{02}	.468 (.139)
S 74	α_{11}	-.033 (.124)		β_1	-.005 (.011) (summer)
S 75	α_{12}	-.000 (.123)			
S 76	α_{13}	-.364 (.123)			
S 77	α_{14}	.045 (.123)			
S 78	α_{15}	-.095 (.123)			
S 79	α_{16}	.143 (.123)			
S 80	α_{17}	.006 (.123)			
S 81	α_{18}	-.177 (.123)			
S 82	α_{19}	-.120 (.123)			
W 73/74	α_{21}	-.206 (.152)		β_2	-.048 (.013) (winter)
W 74/75	α_{22}	-.218 (.147)			
W 75/76	α_{23}	-.113 (.150)			
W 76/77	α_{24}	-.287 (.150)			
W 77/78	α_{25}	-.282 (.147)			
W 78/79	α_{26}	-.354 (.146)			
W 79/80	α_{27}	-.382 (.148)			
W 80/81	α_{28}	-.577 (.151)			
W 81/82	α_{29}	-.418 (.148)			
	γ	-.611 (.098)		γ	-.580 (.098)
	ω	.721 (.083)		ω	.705 (.090)

Figure 5.1(a)

Plot of estimated yearly effects of LCO_t from model (5.5) (winter season)

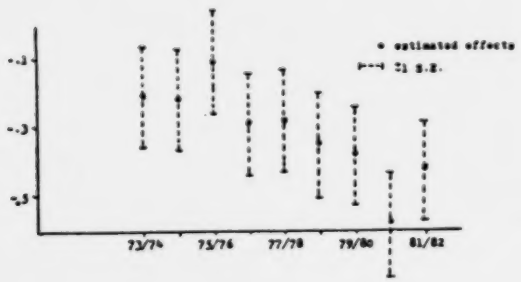


Figure 5.1(b)

Plot of estimated yearly effects of LCO_t from model (5.5) (summer season)

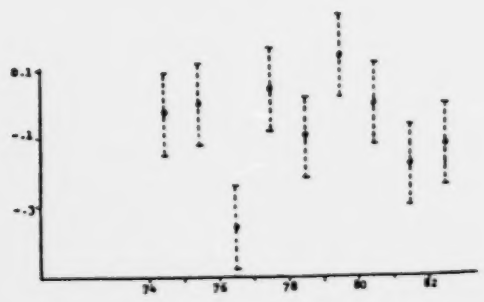


Table 5.2

Estimation Results for the Time Series Models (5.5) and (5.6)

$Y_t = LCO_t^*$ (10/72 - 9/82)

	Model (5.5)			Model (5.6)	
	β_{01}	3.140 (.366)		β_{01}	2.927 (.360)
	β_{02}	.508 (.159)		β_{02}	.525 (.140)
S 74	α_{11}	-.092 (.123)		β_1	-.023 (.011) (summer)
S 75	α_{12}	-.106 (.123)			
S 76	α_{13}	-.499 (.123)			
S 77	α_{14}	-.109 (.123)			
S 78	α_{15}	-.261 (.123)			
S 79	α_{16}	-.031 (.123)			
S 80	α_{17}	-.175 (.123)			
S 81	α_{18}	-.361 (.123)			
S 82	α_{19}	-.307 (.123)			
W 73/74	α_{21}	-.277 (.152)		β_2	-.071 (.013) (winter)
W 74/75	α_{22}	-.342 (.146)			
W 75/76	α_{23}	-.278 (.149)			
W 76/77	α_{24}	-.477 (.150)			
W 77/78	α_{25}	-.487 (.147)			
W 78/79	α_{26}	-.570 (.145)			
W 79/80	α_{27}	-.606 (.148)			
W 80/81	α_{28}	-.807 (.150)			
W 81/82	α_{29}	-.650 (.147)			
	γ	-.614 (.098)		γ	-.574 (.099)
	ω	.676 (.083)		ω	.655 (.090)

Figure 5.2(a)

Plot of estimated yearly effects of LCO_t^* from model (5.5) (winter season)

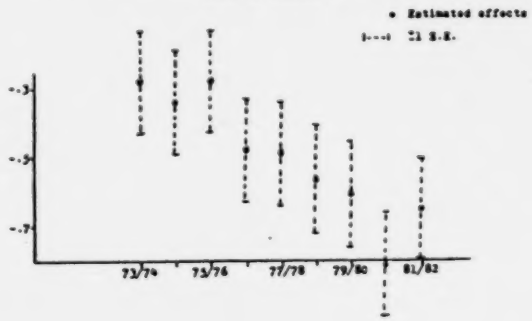


Figure 5.2(b)

Plot of estimated yearly effects of LCO_t^* from model (5.5) (summer Season)

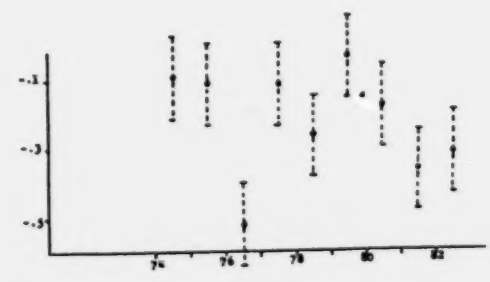


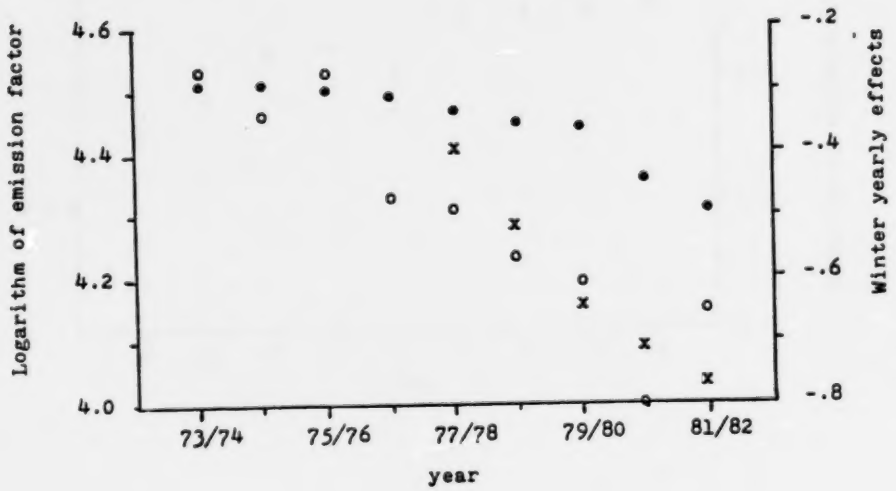
Table 6.1

CO Emission Factors (g/mile) Derived From MOBILE3 Analysis
(Phoenix Central, 7-10 am and 7-10 pm in Oct.-Feb.)

Year & month	No I/M	with I/M
72/1	88.08	
73/1	89.06	
74/1	91.25	
75/1	91.43	
76/1	90.01	
77/1	88.96	
78/1	87.35	81.83
79/1	85.51	72.32
80/1	81.75	63.31
81/1	78.23	59.89
82/1	73.94	56.34
83/1	69.56	52.78

Figure 6.1(a)

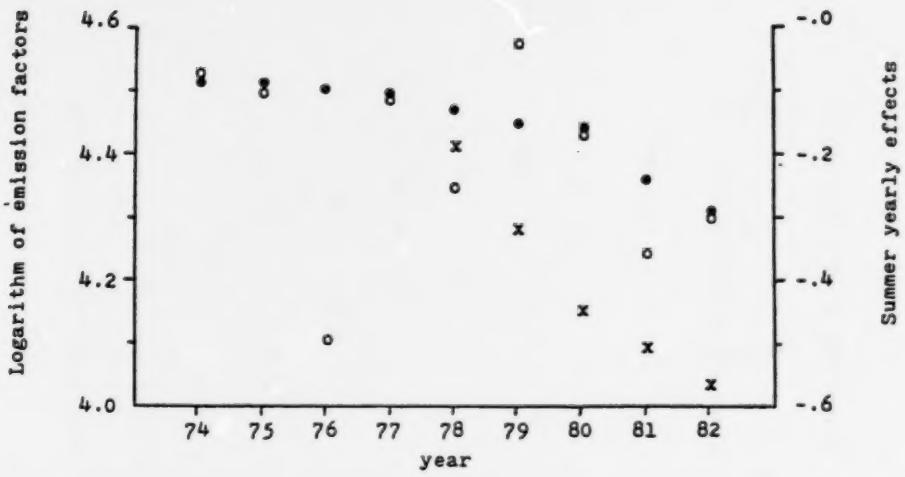
Plot of CO emission factors with yearly effects in the winter season



- • Emission factors without I/M (based on MOBILE3)
- x Emission factors with I/M (based on MOBILE3)
- o Winter yearly effect based on Model 5.5

Figure 6.1(b)

Plot of CO emission factors with yearly effects in the summer season



- • Emission factors without I/M (based on MOBILE3)
- x Emission factors with I/M (based on MOBILE3)
- o Summer yearly effect based on Model 5.5

SAMPLING DESIGN: SOME VERY PRACTICAL CONSIDERATIONS

D. E. Splitstone
IT Corporation, 2790 Mosside Blvd., Monroeville, PA 15146

INTRODUCTION

When Walter Liggett extended the invitation to participate in this conference, I glibly responded in the affirmative and in a cavalier manner conceived of a topic title that, I perceived, would not get me in too much trouble. I will confess the possibility of following George Tiao gave me some pause. However, I am about to provide some words of what I hope will be wisdom regarding very practical considerations in sampling design for the gathering of information during studies of the environment.

Because my current position of employment requires a concentration on the problems posed by hazardous materials and wastes, this paper will identify a few of the more intriguing statistical issues associated with hazardous waste investigations. As my approach to these issues is conditioned on my subscription to Donald Marquardt's wider view of the statistical profession, permit me to quote his definition of that perspective in order to provide a frame of reference.

The wider view is that statisticians are systems guides for tough, fuzzy issues wherever the collection and interpretation of data should be involved. The wider view expects statisticians to provide the following:

- Full service guidance and participation in problem diagnosis and solution
- Services for problems in any area of application
- Services at any level of detail or abstraction.

From this perspective, individual statistical methods are viewed as tools or component elements in an overall systems approach (Marquardt, 1987).

Philisophically, the design of any sampling scheme is easy. Paraphrasing William Lurie's classic article (Lurie, 1958) one only needs to answer the following questions:

- What does one know about the site under investigation?
- What does one wish to determine regarding the site?
- And, how certain does one wish to be in the conclusions?

In addition, one may add the admonition of Oscar Kempthorne (ca. 1966) that an objective of any experimental design is to make the analysis of the resulting data easy.

The practicality of meeting of all of these criteria, however, is quite another matter.

EXAMPLE 1 - THE CONTAMINATED BUILDING

Consider as an illustrative example the situation presented by a building thought to be contaminated by polychlorinated biphenyls (PCBs), (Ubinger, 1987). The design of a sampling program for such a situation is no trivial task. It certainly must start with a clear statement of the objectives which may be multifaceted. Consider the following possibilities:

- Determination of the existence of contamination (a preliminary site investigation)
- An assessment of the risk of exposure
- Tracking the spread of contamination and mapping its extent
- Identification of sources of contamination
- Post cleanup confirmatory sampling, and
- Certification that cleanup has been achieved (closure sampling).

There is not single sampling strategy that will apply to all types of problems and all of the above objectives. Clearly, the objective of any sampling strategy is to provide the most cost-effective approach to meeting the study objectives while maintaining a high degree of confidence, precision and accuracy regarding the study's conclusions.

Once the objectives of the investigation are clearly defined, the approach to the development of a sampling strategy must consider the nature of the source of contamination. The two major causes of PCB contamination are electrical transformer fires and historical usage of PCB-containing lubricants. Each results in very different patterns of potential contamination which may affect the design of data collection. Typical patterns are as follows:

Fires/explosions

- Fluid loss - Pattern radiating from the source; largely confined to horizontal surfaces at or below the source
- Smoke borne - Spread through ventilation system, decreasing contamination as distance from the source increases; horizontal and vertical surfaces at and above the level of the source

- Tracking by human activity - Floors and high-contact surfaces; may see buildup in high traffic areas

Historical Usage

- Oil creep - Generally on horizontal surfaces in proximity to the source; may be localized buildup in less accessible areas
- Reclamation and storage rooms and facilities
- Tracking - floor and other high-contact areas
- Other - permeation of floor into underlying soil; handling areas - loading docks, trash areas.

Let us assume for the moment that the initial objective of the investigation is to define the extent of contamination in anticipation of building cleanup or remediation. Traditionally, the sampling scheme employed in such situations has been to engage in directed or judgemental sampling. That is, sample those areas which appear through evidence of oily stain to be contaminated, and occasionally sample areas which do not visually appear to be contaminated. The statistical effect of directed sampling is certainly to increase the precision of the estimated magnitude of contamination in those areas which visually appear contaminated, but is of little assistance in defining with confidence the spatial extent of contamination. Nor does directed sampling provide sufficient information on the lack of contamination, or the cleanliness of areas which appear to be uncontaminated so that unnecessary remediation can be avoided with confidence (Ubinger, 1987).

Though I will return to the issue of making decisions regarding cleanliness later, let me simply state that a decision to require remediation can be made based on only one sample. The decision not to remediate, however, requires much more evidence. Thus, if an objective is to minimize unnecessary remediation, equal or perhaps greater attention in sampling design should be given to apparently "clean" areas as was accorded apparently contaminated areas.

Ubinger (1987) has compared the results of statistically designed sampling and directed sampling and found the two methods to be equally effective in detecting areas of high contamination with an obvious advantage to the statistically designed sampling program of providing unbiased results. In addition, the statistically designed program was more effective in supporting assessment of the spatial extent of contamination and in identifying areas which are uncontaminated.

It appears obvious that there are advantages to statistically designed sampling plans. But what type of statistical design does one employ to do an effective job? The answer to that question can only be that it depends. It depends on the specific objectives of the investigation and what is known about the site and potential source of contamination. One generally starts by laying a grid system over the area of interest. The grid spacing is determined in large measure from decisions regarding the areal size and geometry of the minimum contaminated area necessary to detect and the degree of risk deemed acceptable of not being able to detect that area (Zirchky and Gilbert, 1984). These decisions regarding target size and acceptable risk are not inherently statistical decisions. They are, appropriately, the policy decisions of those who must balance cost and risk. There may not be a single decision regarding target size and the risk of missing that target applicable to the whole of the area or building. For instance, if the objective is to assess the health risk to building occupants due to PCB exposure, areas with high human contact potential are of greater importance than areas of low contact potential. Thus, it may be perfectly justifiable to engage in less sampling in low contact areas than in high contact areas.

It would certainly appear that stratification is an important component in the sampling design. Intelligent choice of strata can only be accomplished through thorough knowledge of the physical plant of the building under construction, the cause of the alleged contamination, and traffic and work patterns within the facility. The latter is less important if the objective of the study is solely to map the spatial extent of contamination; however, if health risk is to be an important component in the decision making process, a probability quota sampling design may be employed. Only if sampling locations are chosen with probability proportional to frequency of contact can an overall estimate of exposure be formed without application of external weights.

In summary, the design of sampling schemes for the allegedly contaminated building are among the most complex to formulate. In order to define the problem statistically, a number of definitions are required:

- What is to be measured; what is a suitable limit of detection?
- What is the criteria for cleanup decisions - are these criteria to be based on risk assessment?
- What are the statistical quantities that define a decision rule for when to remove material (the specific value may be a result of a site-specific risk assessment), and the quantification of the risk of failing to remediate?
- What field sampling plan is to be used to obtain <u>representative</u> PCB concentration data?

- What are the action guidelines?
- If the decision rule says "clean," must the whole room be cleaned?
- If only hot spots are found, might they alone be cleaned?
- Is there a logical minimum to the size of a "remediation unit?"

Obviously, most of the answers to these questions are site specific and, therefore, defy general codification. The answers also require the direct interaction of the statistician with other members of the study team, the interchange of information among the expert in building ventilation, the industrial engineer, the expert in building remediation, and the risk assessment specialist.

EXAMPLE 2 - THE IMPOUNDMENT

Compared to the problems associated with the sampling of building interiors, the sampling design for a waste impoundment might appear trivial. One does not have to reflect on the sampling of walls, ceilings, air handling systems and how to deal with lighting fixtures and machinery. However, first impressions are often wrong, and though certainly easier to conceptualize, the sampling design for a waste impoundment is not trivial. Returning to the central theme of this paper, one must start by asking questions.

- Just what does one wish to find out about the impoundment and its contents?
- What types of waste might the impoundment contain and at what levels of concentration are they expected to be found?
- How did the waste get there? By pipe, from a process reasonably continuous in time or by truck from some batch process?
- Where are the discharge points to the impoundment?

Answers to questions such as these provide the clues to potential spatial stratification for which the sampling design must account.

Generally, whether the waste within the impoundment is spatially homogeneous or heterogeneous is of concern regardless of what will be done with it later. This applies largely to sludges, as any liquid layer will likely be well mixed at the moment of treatment and disposal, except perhaps in very large impoundments. Considering the sludge alone, prior to the design of a sampling strategy, the potential stabilization must be considered. If no stabilization is to occur, then the sampling design must focus on the chemical characterization of the sludge in situ subject to the conditions of closure or removal. If the sludge is to be removed, the process of removal may well suggest natural sampling units, e.g., a volume associated with a backhoe or clam shell bucket or perhaps a truck load. The delineation of subareas containing hazardous or nonhazardous material may be useful in reducing the cost of treatment and/or disposal. If stabilization is to occur, the sludge may be homogenized vertically by a barge.

Popular guidance currently provided for the construction of sampling plans for the characterization of potentially hazardous waste (U.S. EPA, SW-846), contains a philosophy which I believe is misleading and statistically naive. Consider the following statement from SW-846.

> The regulations pertaining to the management of hazardous wastes contain three references regarding the sampling of solid wastes for analytical properties. The first reference, which occurs throughout the regulations, requires that representative samples of waste be collected and defines representative samples as exhibiting average properties of the whole waste.

This statement presumes two things which are almost never true: (1) that the waste is homogeneous, and (2) that environmental protection is served by considering what happens on the average. It has been my experience that hazardous waste is almost always heterogeneous in space and/or time except under very special circumstances. It certainly is appropriate to require that samples be representative of the portion of the waste from which they were collected, but to define "...representative samples as exhibiting average properties of the whole waste" expresses a view of the world that simply is not true. Homogeneity is something which must be determined. Appropriate situation specific sampling designs can effectively be used for this purpose. How effective and efficient the sampling design is for characterizing a hazardous waste population depends in large measure upon the amount of prior knowledge available about the specific situation and a statement as to the objective uses of the data collected.

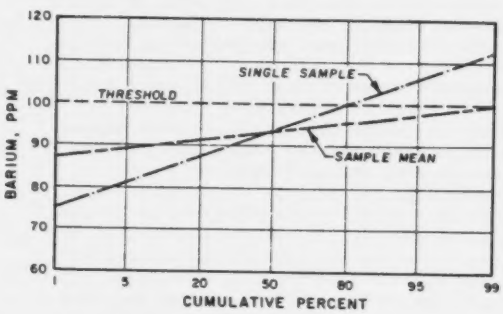
The fact that what happens on the average rarely serves the purpose of environmental protection seems obvious to me, and I will later attempt to demonstrate this very point. I submit that it is not what happens on the average but the consequences of the extremes that are environmentally interesting. As will be illustrated later, when discussing the example given in SW-846, it gives one little comfort to know that in expectation the average of nine samples of sludge will be less than the specified regulatory threshold 80 percent of the time when a full 20 percent of the individual samples will exceed that threshold.

The second paragraph of Section 1.1.1 indicates that the issue addressed by representative samples is that of sampling accuracy. I submit that this is not the case under the definition of representative samples stated. Further, this may not be the case for the subpopulation of waste from which the sample was taken unless use of various analytical methods and interlaboratory comparisons are employed.

> The second reference, which pertains just to petitions to exclude wastes from being listed as hazardous wastes, specifies that enough samples (but in no case less than four samples) be collected over a period of time sufficient to represent the variability of the wastes.

The variability of the waste is certainly important to assess, however, rarely is the variability exhibited among samples taken during even a well designed investigation due solely to random variations in the waste. Much of the variability observed can be attributed to factors in space and/or time. Ignoring such factors can have significant consequences, not only in terms of the degree of environmental protection, but also in terms of unwarranted expense and the sensibility of statistical tests employed to evaluate the waste. Sampling designs which permit identification of these factors allow for the possibility of increased environmental protection at lower cost. Unfortunately, the identification of appropriate subpopulations may require the application of statistical techniques in light of specific knowledge of the situation in a manner which cannot be codified. In any event, the number and placement of samples in space and/or time for hazardous waste investigations has many more aspects than one of sampling precision once one conceives of the problem more realistically than that of characterizing average properties.

Consider the following illustration of the consequences of basing the decision as to whether a waste is hazardous or not hazardous on only average properties. The figure below is generated from data presented as the Hypothetical Example given on Page 14 of SW-846 and used to illustrate sample size selection under different strategies. In construction of the figure, I have assumed that the normal distribution is an adequate model and that the true mean (μ) is equal to 93.56, and the true standard deviation is equal to 7.75. These values are the estimates of their respective parameters at the moment that a decision is made that barium is not present in the sludge at a hazardous level (see Equation 6 on Page 17).



DISTRIBUTION OF SINGLE SAMPLES AND SAMPLE MEANS
BARIUM SLUDGE EXAMPLE SW-846

The salient points to be made using this figure are simply stated in the following: (1) Though a conclusion was reached that barium was not present at a hazardous level in the sludge based on inferences regarding the mean of nine samples, a full 20 percent of the sludge would be expected to be above the hazardous level. And, (2) it would appear that if the true mean concentration is below the regulatory threshold, a decision as to whether a constituent is or not present at a hazardous level is solely a function of the number of samples taken using the logic presented.

In point 1 above, I have used the term expected in the statistical sense. If one were to demand a greater degree of confidence in estimating the maximum percentage above the regulatory threshold, then a tolerance limit rather than a confidence limit should be used. The result would be closer to 30 percent of the sludge in the hazardous concentration range with 95 percent confidence. Guttman, 1970, differentiates these approaches as "Beta content" versus "Beta expectation" tolerance regions.

If one focuses on deciding whether or not the sludge contents of an impoundment are hazardous, given a definition of "hazardous," a specification of risk to be tolerated in terms of the fraction of sludge permitted to exceed the defined threshold, and a specified desired degree of confidence to be associated with the decision, then determination of the number of samples required would appear straightforward. However, many would attempt such a determination by using techniques which assume that the statistical distribution of chemical concentrations is normal. Though upon occasion, one may be willing to assume that variations introduced by measurement and sampling may arise from a normal distribution, in general the assumption of a normal distribution of hazardous waste measurements over space and/or time cannot be supported. Morse, et al, (1986) suggest that nonparametric approach to the problem be taken. In such an approach, the maximum proportion of the population exceeding the ultimate or penultimate observation with specified degree of confidence is a function of the number of samples taken. For instance, one is 95 percent confident that at most five percent of all samples will have a concentration greater than the maximum concentration of a set of 59 random samples.

The use of order statistics and the rank orders of hazardous waste chemical concentration data are quite useful in hazardous waste investigations. However, that is a topic for an in-depth presentation of data analysis. Such data analysis considerations do indeed have an effect on the sampling design.

Returning to the problem of making a hazardous/nonhazardous declaration regarding the sludge in an impoundment, a first step might be to consider it as a target detection problem. The size and geometry of the target is determined by one's knowledge of how the waste entered the pond. Entry in the form of a slurry through a pipe at one end of the impoundment might lead one to consider a rather large elliptical target, for instance. A grid may then be superimposed on the pond surface, which will give a high probability of detecting the predefined target if samples are collected at, say, grid nodes. Quite likely, the cost of analyzing all of the collected samples will be quite high. However, if the concentration decision level is sufficiently large, sample compositing may be effectively introduced to keep the costs within reason. Of course, the assumption here is that the analytical cost is large with respect to the sampling cost. This assumption is not always true.

Consider for example an impoundment of nominal dimensions of 200 by 150 feet. Superimposing a square sampling grid with 25-foot grid spacing gives one a probability of 80 percent of detecting a circular area of contamination of 25 feet diameter with 48 samples. If one is concerned about contamination with barium, and the concentration level for a declaration of a hazard is 100 ppm, then the possibility of compositing samples in sets of four will reduce the number of analyses to 12 without sacrifice to environmental protection, provided that the analytical method detection limit is less than 25 ppm. One must consider, however, the potential losses associated with an increase in Type I error should the sum of the concentrations of the four individual samples, now unknown, exceed 100 ppm when each sample concentration is in fact less than 100 ppm. This increased risk of unnecessary remediation may be reduced if sample holding times are such that aliquots of the individual samples can be preserved for later analysis. If such a two-stage analytical program is not possible, then some tough risk balancing decisions must be made prior to the design of the sampling program.

As a final comment regarding the waste impoundment, if stabilization is to take place, a two-stage sampling program may be the most cost effective. The purpose of the first stage of sampling may be solely to determine the gross trends in the physical/chemical properties of the sludge. Interest is primarily in determining those factors which may affect stabilization. At this stage, only a few samples are required to investigate gross trends, say only nine for the impoundment described above. Multiple regression techniques may be used to investigate the existence of spatial trends important to cost-effective application of stabilizing treatments. Once stabilization has been accomplished, the objective of additional sampling would be to determine the requirements of sludge disposal or impoundment closure. As the act of stabilization almost invariably involves mixing, the properties of the stabilized sludge may be quite different than the unstabilized material.

EXAMPLE 3 - THE LANDFILL AND THE INDUSTRIAL OR RESIDENTIAL SITE

Much of the above discussion applies to the investigation of landfills or the potentially contaminated industrial site. A very important difference between this type of site and the waste impoundment is the greater potential for lack of spatial homogeneity. Obviously, in the instance of a modern, well-run and documented landfill, the sampling plan should be designed to comprehend the known stratification of material disposition. However, consider the situation of the "Superfund" site. Little information is generally available as to the depositional location of potentially hazardous material. Certainly a continuum of the degree of information exists regarding the initial deposition, potential transport and nature of the waste from well kept documentation, to reliance on memory and to nothing at all. The same is true of the industrial site where contamination may have occurred at some past time through, not only intentional disposal, but accidental spills or process leaks. One must also be cognizant of construction activities and their potential for transportation of contaminated soil. The more of this information that is available, the better the sampling plan.

As is the case with the waste impoundment, it seems reasonable to begin the design of a sampling plan by considering the problem as a target detection problem. Once the possible mechanisms for hazardous material deposition have been identified, the size of the target to be detected and the degree of risk to be tolerated in not detecting the target must be determined. The latter is purely a policy decision, but the former may be assessed in more technical terms. Suppose it is known that organic solvents were disposed of in 55-gallon drums. Given some information regarding the soil permeability, the transport coefficient of the solvent and the approximate length of time a drum has been in the ground, it is possible to estimate the spatial extent of a potential plume arising from a leaking drum. However, the task is more complex in the case of dioxin which is generally not mobile in soil. In this case, very small isolated deposits of dioxin may exist. This is particularly true on the industrial chemical plant site where deposition may have been a result of explosion, process leakage or plant expansion activities.

If the analytical procedure to be employed is inexpensive, as with a walk-over radiation survey using a geiger counter or a portable organic vapor analyzer, one may simply lay out the required grid network for the desired target size and risk of not detecting the target. Such situations are rare, however. Most frequently the statistician's job is to effect a compromise between the competing objectives of determining the potential of hazard and minimizing the cost of doing so.

As indicated in the above discussion of the contaminated building, it takes only one sample to indicate the existence of contamination. It takes several more samples to determine the extent of contamination however. Often contamination is known to exist at a site through a prior sampling event. Possibly the location of a contaminated area is even known quite precisely. The objective of future study, and hence the sampling plan, is to determine the extent of contamination and provide a rational basis for the making of remedial decisions. If one is willing to make the assumption of the stationarity of spatial similarity of contaminant concentrations, then the techniques used in geostatistics offer some exciting possibilities. This is particularly true of nonparametric geostatistics (Flatman 1984: Isaaks, 1984; Journal 1984-1, 1984-b, 1987). In such instances one may be willing to employ a sampling grid of rather course spacing, augmenting it at an occasional grid node by sampling along a geologic "fence."

There is no substitute for knowledge. To be effective in meeting the objectives of any investigation, the sampling scheme must be designed with all the available knowledge regarding the site, the suspected contaminant, its potential for transport, its means of initial deposition and potential analytical methods required for its detection. This requires a close interaction among all members of the team of investigators. Only with site-specific knowledge can an adequate sampling plan be developed.

Perhaps the most difficult objective to accomplish is to declare with a specified degree of certainty, that a site is free of contamination. One is faced with this objective after remediation of a contaminated site and under the provisions of the State of New Jersey Environmental Cleanup Responsibility Act (ECRA). ECRA is involved in all major real estate transfers in New Jersey, and similar legislation is pending in several other states.

Exner, Gilbert and Kinnison (1984) have taken a novel approach to deciding when remediation is complete for dioxin contaminated sites near Times Beach, Missouri. Their proposed strategy divides the known area of contamination into "cleanup units" of a size conducive to the use of appropriate soil removal equipment (e.g., large earth moving equipment). A practical unit size is 20 by 250 feet. Adjacent clean-up units are established to ring the unit in which remediation is initiated. A grid is established for each unit by the intersection of two lines parallel to the long axis of the unit, spaced ten feet apart starting five feet from one end of the unit. A fixed geometric sampling pattern is established with each grid square. Samples from the same position in each square are then composited. A number of aliquots, n, are then selected randomly from each composite for analysis. If there are m composite samples, then there are N=nm data for each clean-up unit.

The N data are then used to calculate the arithmetic mean and standard deviation of the n composite means. These statistics are then used to compute an upper confidence limit for the true mean concentration for the clean-up unit. The degree of confidence is consonant with the risk of not remediating a contaminated unit when the environmental risk is based on the mean unit concentration. If this upper confidence limit statistic is above a decision criteria (an acceptable true mean concentration), then another two inches of soil is removed from the entire unit and the test repeated. Otherwise, the unit is deemed to be clean and remediation stops. Because of the manner of statement of the null hypothesis, this procedure controls the risk of not remediating a contaminated unit. The risk of unnecessarily remediating a clean unit remains uncontrolled however. The financial loss of this uncontrolled risk is perhaps quite small in that the cost of removal of an additional two inches of soil from a clean-up unit is small once mobilization has occurred. This is particularly true if there are sufficient clean-up units to keep a field team occupied while awaiting the analytical results from a prior remediated unit test.

If the site under consideration has not previously been declared contaminated, then the situation is much more complex. The complexity derives not from the lack of sufficient statistical techniques with which to design sampling plans, but from the need to balance possible risk with the cost of achieving a specified level of risk control. In part, I believe that this is a cultural problem. We are conditioned to finding something. If we do not find contamination, we feel disappointed. As a society, we must come to grips with defining the levels of risk we are willing to take and be willing to bear the cost of achieving those levels. Organizations who find themselves subject to legislation such as ECRA, must balance the risk of not finding contamination, when in fact, it exists with the cost of the potential liability in the future. The statistician's professional role is confined to the definition of risk alternatives and providing sampling designs consonant with the policy decisions made in response to those definitions.

EXAMPLE 4 - GROUND WATER MONITORING

Rarely has a specification of statistical procedures created as much controversy as in the case of the monitoring of ground water around hazardous waste facilities. Since the promulgation of the regulations now found in 40 CFR 265.92, there has been repeated criticism of the application of Student's "t" test in any form to the determination of ground water contamination by a hazardous waste facility. Under federal regulations (40 CFR 265 and 264), the owners and/or operators of hazardous waste facilities are required to monitor the ground water around the hazardous waste facility and to detect the existence, if any, of ground water contamination. While this appears a perfectly reasonable objective, the procedures prescribed for doing so are, in most cases, at best naive and at worst wrong (see Liggett, 1985 and Splitstone, 1986). Here is a clear case where the promulgation of a statistical hypothesis testing procedure has relegated the importance of statistical design to a secondary position rather than the other way around.

The objective of monitoring ground water near a hazardous waste facility is clear. It is to determine whether the facility is contaminating the ground water. A sampling program to make such a determination should be designed so that any contribution from the particular facility can be distinguished from other sources. These other sources, which may be both anthropogenic and natural, provide what is loosely referred to as the background. The design of a sampling program to distinguish the contribution of the site interest from the background is complicated by temporal and spatial variations in the background.

Temporal variations in ground water quality may in part be explained by recharge rates which are weather related. Spatial variations among wells may be due to a variety of reasons. Some of this variability may be reduced by ensuring that the wells monitoring the aquifer beneath a particular waste disposal site are located relatively close together and drilled to the same depth. Nevertheless, spatial variations may occur even if there is no contribution from the site being monitored. These variations might be expected to be a result of horizontal gradients across the site, unexpected confluence of different aquifers, or more likely, the natural geochemical conditions of the area.

If one is fortunate enough to be able to begin monitoring prior to the existence of the facility, life is quite easy. It is then possible to investigate the existence of potential temporal and spatial trends in ground water quality without fear of any confounding from the facility. Such a circumstance, however, is rare indeed. More likely, monitoring begins around an existing facility. In either case, it would appear advisable to design a study to investigate the characteristics inherent to ground water quality in the environs of the site or proposed site prior to designing a long-term monitoring program.

The design of the initial investigation, where it includes the siting of new wells or simply the analysis of existing data begins by asking what is known or suspected about the site. Input from the hydrogeologist and geochemist is a must. Knowledge regarding the probable direction of ground water flow, possible seasonal or diurnal fluctuations and the potential chemical changes in ground water quality as it moves through the local rock and soil is essential to the intelligent interpretation of the data. The result of a well-designed initial investigation should be a "model" describing natural or "background" variations in ground water quality. Such a model is of necessity site specific. The design of a long-term ground water quality monitoring program should have as its objective the detection of deviations from this background model. Techniques appropriate to testing hypotheses regarding deviations from this background model must be employed and should be permitted by federal regulations.

CONCLUDING REMARKS

The design of schemes to collect data for the purpose of providing meaningful information regarding potential hazardous waste sites is of necessity site specific. The design must be based on a clear statement of the information gathering objective and can only be efficient if careful consideration is given to the accumulated prior knowledge regarding the site.

Attempts that have been made to specify the statistical techniques to be employed in environmental decision making often ignore the presence of site-specific conditions which may have a dramatic effect on the outcome of the prescribed mathematical calculations. Realizing that it is not possible to promulgate regulations applicable to every site, a clear regulatory statement of the objective to be achieved coupled with a specification of the degree of certainty of achieving that objective would appear to be a reasonable regulatory goal. The details of data collection and interpretation are appropriately left to the statistician in concert with the knowledgeable scientists and engineers.

The design of a sampling strategy for any site investigation is more a logical thought process that the application of statistical tools. The purpose of environmental regulation might well be served by carefully outlining the steps in this thought process and schooling those charged with the protection of the environment in their application. The role of statistics and the statistician in this process have been outlined nicely in the published notes from a short course on nuclear site decontamination and decommissioning (Barnes, 1981).

REFERENCES

Barnes, M. G., 1981, "Statistics and the Statistician in Nuclear Site Decontamination and Decommissioning - Lecture Notes for a 4-Day Short Course," U.S. DOE Contract DE-AC06-76RLO 1830, Pacific Northwest Laboratory, Richland, West Virginia.

Exner, J. N., R. O. Gilbert and R. R. Kinnison, 1984, "A Sampling Strategy for Cleanup of Dioxin in Soil," Appendix I, Quail Run Site Hazard Mitigation Plan, Environmental Emergency Services Company, Chesterfield, Missouri.

Flatman G., 1984, "Using Geostatistics in Assessing Lead Contamination Near Smelters," Environmental Sampling for Hazardous Waste Sites, ed. Schweitzer, ACS, Washington, D.C.

Guttman, I., 1970, "Statistical Tolerance Regions: Classical and Bayesian," Griffin's Statistical Monographs and Courses, No. 26, Hafner Publishing Co., Darien, Connecticut.

Isaaks, E.H., 1984 Risk Qualified Mappings for Hazardous Waste. A Case Study in Non-parametric Geostatistics, MSc Thesis, Branner Earth Sciences Library, Stanford Univ.

Journel, A.G., 1984-a, "The place of non-parametric Geostatistics, "Geostatistics for Natural Resources Characterization, ed. Verly et al., Reidel, Dordrecht, pp. 307-35.

Journel, A.G., 1984-b, "Indicator Approach to Toxic Chemical Sites, "EPA, EMSL-Las Vegas, Report of Project No. CR-811235-02-0.

Journel, A. G., 1987, Short Course, "Advanced Geostatistics For Mining and the Environment," sponsored by Stanford University, and U.S. EPA Las Vegas, January 12-16, Las Vegas, Nevada.

Kempthorne, O., ca 1966, Course Notes "Design of Experiments," Department of Statistics, Iowa State University, Ames, Iowa.

Liggett, W., 1985, "Statistical Aspects of Designs for Studying of Contamination," Quality Assurance for Environmental Measurement, ASTM STP867.

Lurie, W., 1958, "The Impertinent Questioner: The Scientist's Guide to the Statistician's Mind," American Scientist, March, pp. 57-61.

Marquardt, Donald W., 1987, "The Importance of Statisticians," Journal of the American Statistical Association, Vol. 82, No. 397, pp. 1-7.

Morse, M. W., Sproat and J. Warren, 1986, "Sampling and Analysis for Delisting Petitions/Statistical Basis for Sampling," Second Annual Symposium on Solid Waste Testing and Quality Assurance, U.S. EPA Conference, July 15-18.

Splitstone, D. E., 1986, "A Statistician's View of Ground Water Monitoring," Proceedings of the National Conference on Hazardous Wastes and Hazardous Materials, Hazardous Materials Control Research Institute, pp. 8-12.

Ubinger, E. B., 1987, "Statistically Valid Sampling Strategies for PCB Contamination," to be presented at the Electric Power Research Institute Seminar on PCB Contamination, Kansas City, Missouri.

U.S. Environmental Protection Agency, Interim Status Standards for Owners and Operators of Hazardous Waste Facilities, 40 CFR 165.92.

U.S. Environmental Protection Agency, Regulations for Owners and Operators of Permitted Hazardous Waste Facilities, 40 CFR 264.97.

U.S. Environmental Protection Agency, 1982, Test Methods for Evaluating Solid Wastes, SW-846, Second Edition, July.

Zirchsky, J. and R. O. Gilbert, 1984, "Detecting Hot Spots at Hazardous Waste Sites," Chemical Engineering, July 9.

DISCUSSION

W. Barnes Johnson
U.S. Environmental Protection Agency, 401 M St. SW, Washington, DC 20460

I have been asked to provide a discussion from the perspective of an Environmental Protection Agency (EPA) statistician. In communicating the EPA perspective it is first important to note that statistics are used in a variety of EPA's program activities. A convenient taxonomy for categorizing those activities begins with the most broad and wide ranging and continues to the most detailed:

RESEARCH AND DEVELOPMENT
POLICY DECISIONS
REGULATORY DEVELOPMENT
GUIDANCE DEVELOPMENT
PERMITTING
COMPLIANCE.

Research and development activities are the most broad and wide ranging in terms of their possible application and impact on EPA's programs. Some of Dr. Tiao's work has been done to support EPA activities at the broad end of the continuum. Policy decisions, regulatory actions, and guidance development can be quite broad, but can also be exacting and specific in terms of application; therefore they reside in the middle of the continuum of areas where statistics might be used at EPA. At the narrow end of the continuum the concerns focus on site specific sampling, analysis, and decision making problems related to permitting and compliance. Dr. Splitstone's experiences and certain aspects of Dr. Tiao's work have been directed toward the problems that are faced at the narrow end of the continuum. The spectrum offered by these two presentations seem to cover the broad and difficult responsibilities that EPA faces. There is a responsibility imposed on EPA to be general so that a national scope is provided; but EPA also has a responsibility to provide technical direction that can be useful at specific locations with unique circumstances.

One thing that becomes clear is that at the EPA statistics are not a product. This is not so true, for example, at the Census Bureau or the Bureau of Labor Statistics; at these Agencies statistics are truly the product of interest. Instead, EPA's products are primarily regulatory decisions. These decisions are based upon environmental, engineering, economic, political, financial, risk, and health analyses. This difference reflects itself in the way that statistics and therefore statisticians are used at EPA. One might even expand this to say that there is a difference in the way problems are characterized, approached, and solved. It is important to remember that, when statistical work is called upon to provide perspective, insight, or response relative to an EPA interest or initiative, the statistical analysis is only a tool that is used in the ultimate decision. For example, in the case of Dr. Splitstone's PCB contamination problem, EPA may have been interested in knowing if the building was a health threat. Knowing that there is a high confidence that no circular hot spots of a reasonably small size are present with PCB's above an action level may or may not be sufficient information for the decision maker. The implications regarding exposure likelihood and the resulting risk potential may need to be analyzed. Similarly, in the case of Dr. Tiao's work, EPA might be interested in deciding whether vehicle inspection and maintenance programs effectively reduce ozone. However, the study may not be useful for making a national policy change given the messy and confounded information from a few sampling stations located in Oregon and Arizona. The primary conclusion is that statistical methods are tools that support EPA's program functions and, as such, they must be designed to allow for or assist in formulating defensible decisions.

Although statistical methods are a tool that can be called upon when an investigator or policy analyst desires, they are not always considered early, during the design phase of a project. Instead, "case study" or ad hoc "data base" approaches are often used as methods for solving complex environmental problems. Consider the following hypothetical examples that explain why these data acquisition methods might be used. An investigator needs to evaluate the nature, status, or extent of a national environmental problem. The investigator decides that data must be obtained that in some sense "represent" the national situation. One approach is to pick in a purposive manner geographic and/or temporal locations which seem to reflect or parallel the national situation. In some cases, locations or situations may be chosen to provide results that are both extreme and average, and in this way bound the problem. This is the "case study" approach.

Another approach, which certainly has been facilitated by information processing technology, is to attempt to collect every piece of data that has ever been accumulated by any source and store the data in a uniform format on a computer. The idea is that the availability of the information in an accessible form ensures a meaningful "representation" of the situation. This is the "data base" approach.

There is no question that these methods have utility and merit given certain conditions or investigative purposes; however, these approaches can also distort subsequent analysis and evaluations. Also, the conclusions that can be made from analyses performed on information collected in such a manner are limited. Dr. Tiao, in one of his examples, was provided data from two stations (one in Corvallis, Oregon, and one in Portland, Oregon). Someone selected these sites, undoubtedly out of a network of sites, and commissioned Dr. Tiao to conduct a detailed and, I must say, excellent time series evaluation. How was it decided that these two sites were chosen? Certainly, it would have been preferable for a statistician to be involved with the experimental design of the sample station monitoring network. Many issues such as between location variablity, frequency and duration of sampling, and timing of instrument calibrations could have been designed into the study and Dr. Tiao clearly pointed these issues out. Instead, interests in acquiring "representative" information by developing a "data base" and performing a "case study" have made the investigation, although interesting and informative, less than it could have been given some design prior to implementation of the sampling. Kruskal and Mosteller (1979a, 1979b, and 1979c) offer interesting background and discussion regarding the meanings of "representative" sampling as used in a variety of technical and non-technical literature sources.

Dr. Splitstone is in the enviable position of being able to influence the statistical design, sampling, and analysis for the purpose of answering specific questions. Dr. Splitstone's examples are site specific and in response to EPA's permitting and compliance programs described earlier. Dr. Splitstone attempts to make no extrapolation beyond the original scope of the sampling program. Through a well conceived statistical design and analysis, he is able to generate data which properly "represent" and answer the specific, statistical questions at hand.

I would like to return to an earlier idea which couches statistical analysis as a decision maker's tool. I also want to continue with the theme that a well conceived statistical design and sampling program offer the best method for generating data which provide a "representative" look at an environmental problem problem. The EPA manager realizes that decisions are made in a risky environment; there is a chance that wrong decisions will be made. Statistically designed sampling programs allow quantification of the likelihood of a wrong decision. However, ad hoc, poorly developed "data base" and "case study" approaches that do not coincide with the objectives under scrutiny, do not allow quantification of the error or risk and certainly provide a biased "representation" of the problem. As such, these data acquistion approaches at worst prohibit and at best constrain the use of statistical analysis as a decision maker's tool for evaluating risk.

There seem to be two general observations from these presentations. First, involvement early in the design process as in Dr. Splitstone's situation is reflected in results that satisfy the objectives of the study and allow decision making in a risky environment. Second, involvement with the data analyses after data collection has been designed and executed will result in analyses and conclusions that must be well qualified and reflective of the limitations of the original sampling program. Dr. Tiao's clear presentations of the strengths and weaknesses of each example he presented certainly support this observation.

REFERENCES

Kruskal, W. and F. Mosteller, 1979a. Representative Sampling I: Non-Scientific Literature. International Statistical Review. 47: 13-24

Kruskal, W. and F. Mosteller, 1979b. Representative Sampling II: Scientific Literature, Excluding Statistics. International Statistical Review. 47: 111-127

Kruskal, W. and F. Mosteller, 1979c. Representative Sampling III: Statistical Literature. International Statistical Review. 47: 245-265

SPATIAL PREDICTION AND SITE SELECTION

Noel Cressie
Department of Statistics, Iowa State University, Ames, IA 50011

Abstract

This article considers various stochastic models that could be used in the prediction of pollutants at locations where no data are available, based on data taken from a spatial network of monitoring sites. The design problem of selecting those sites is also discussed.

Introduction

The "default" stochastic model for a batch of measurements on a phenomenon of interest, is that they are independent and identically distributed (i.i.d.). But the notion that data close together in time or space are likely to be correlated (i.e., cannot be modeled as independent) is a natural one, and will receive particular attention in this article.

Pure temporal models, or time series models as they have come to be known, are usually based on identically distributed observations that are dependent and occur at equally spaced time points. The ARMA models, popularized by Box and Jenkins (1970), are important examples. Spatial models of the type used in this article have to be more flexible since there is not the analogous ordering in space and it is not reasonable to assume that spatial locations of data occur regularly.

Relatively recently (the last 10 to 15 years), meteorological space-time data sets have been collected for the purpose of studying the effects of atmospheric pollution and, in particular, acid rain (see e.g. Peters and Bonelli, 1982). Daily data collection over a number of years at various locations throughout the Northeast USA yields a massive data set. But most of it is temporal so that spatially speaking the data are still rather sparse. Nevertheless spatial prediction is just as important as temporal prediction, since people living in those cities and rural districts without monitoring stations have the same right to know whether their water or their air is polluted. Thus a map of e.g., wet deposition (acid rain), is needed based on samples collected from a spatial network of monitoring sites.

One could take either a global or a local approach to spatial prediction. Global methods such as trend surface analysis attempt to model the spatial variation of the phenomenon in terms of large structural components (e.g., polynomial trend). However the error process usually contains small-scale spatially correlated variation. Local methods such as kriging attempt to predict unknown values from weighted averages of nearby values.

The spatial dependence can also be exploited in the design phase, to establish an "efficient" network of sites. Or if a given network needs a site added or deleted, where should this occur? I shall address these questions using a stochastic spatial model, but it should not be forgotten that there are important considerations to take into account before statistical optimization is performed. Among these are:

- Determine what to measure and how to measure it.
- Be aware of such preliminary variables as local weather conditions, historical data, and take "baseline" data.
- Will measurements be replicated or not?
- Determine the spatial support (level of aggregation) of the measurements.

Then finally, remember the following aphorism of data collection: The necessary is too expensive and the affordable is inadequate.

2. Spatial Prediction

Suppose that measurements, both actual and potential, are denoted

$$\{ z(\underline{s}): \underline{s} \in D \},$$

where $\underline{s}$ is a spatial location vector in $\mathbb{R}^2$ or $\mathbb{R}^3$ (two- or three-dimensional Euclidean space). The index set D gives the extent of the region of interest. Data

$$\{z(\underline{s}_1),\ldots,z(\underline{s}_n)\} \tag{2.1}$$

are observed at known sites

$$\{\underline{s}_1,\ldots,\underline{s}_n\} . \tag{2.2}$$

Henceforth I shall assume that $\{ z(\underline{s}): \underline{s} \in D \}$ is a realization of a stochastic process

$$\{ Z(\underline{s}): \underline{s} \in D \} . \tag{2.3}$$

But is this a realistic model when it is known that (apart from measurement error) $z(\cdot)$ is a deterministic? What is the source of the randomness in (2.3)? To justify the stochastic models used in this article, I shall argue as a Bayesian would. The random variation comes from the uncertainty caused by not observing $z(\cdot)$ at all locations in D. In fact with no data the "prior" might be that the physical process behaves as a second-order stationary stochastic process would. This is a standard tactic in science; when there are infinitely many possible surfaces $z(\cdot)$, deal with the possibilities statistically. Now once data (2.1) are observed, update the prior and work with the posterior distributions

$$\{ Z(\underline{s}): \underline{s} \in D \mid Z(\underline{s}_1),\ldots,Z(\underline{s}_n) \} .$$

From the data (2.1) it is required to predict (possibly noiseless versions of)

$$Z(\underset{\sim}{s}_0) \text{ or } (1/|B|)\int_B Z(\underset{\sim}{s})d\underset{\sim}{s}, \text{ where } |B| = \int_B d\underset{\sim}{s}.$$

I shall present both the trend surface model and the random field model, and then compare the two.

2.1 Trend surface model

Suppose $Z(\cdot)$ can be written as

$$Z(\underset{\sim}{s}) = \mu_1(\underset{\sim}{s}) + \varepsilon_1(\underset{\sim}{s}); \quad \underset{\sim}{s}\in D, \tag{2.4}$$

where $\varepsilon_1(\cdot)$ is a zero mean white-noise process; i.e.,

$$E(\varepsilon_1(\underset{\sim}{s})) = 0 \; ; \; \underset{\sim}{s}\in D$$

$$\operatorname{cov}(\varepsilon_1(\underset{\sim}{s}), \varepsilon_1(\underset{\sim}{u})) = \begin{cases} \sigma_1^2; & \underset{\sim}{s} = \underset{\sim}{u} \\ 0 \; ; & \underset{\sim}{s} \neq \underset{\sim}{u}; \; \underset{\sim}{s},\underset{\sim}{u}\in D. \end{cases} \tag{2.5}$$

Furthermore, suppose the trend surface $\mu_1(\cdot)$ can be written as,

$$\mu_1(\underset{\sim}{s}) = \beta_1 f_1(\underset{\sim}{s}) + \ldots + \beta_p f_p(\underset{\sim}{s}); \quad \underset{\sim}{s}\in D, \tag{2.6}$$

where $f_1(\underset{\sim}{s}),\ldots,f_p(\underset{\sim}{s})$ are known functions of spatial location $\underset{\sim}{s}$. Notate the unknown coefficients as

$$\underset{\sim}{\beta} = (\beta_1,\ldots,\beta_p)', \tag{2.7}$$

the data as,

$$\underset{\sim}{Z} = (Z(\underset{\sim}{s}_1),\ldots,Z(\underset{\sim}{s}_n))', \tag{2.8}$$

the errors as

$$\underset{\sim}{\varepsilon} = (\varepsilon(s_1),\ldots,\varepsilon(\underset{\sim}{s}_n))', \tag{2.9}$$

and the trend-surface components as,

$$F = \{f_1(\underset{\sim}{s}_1) \ldots f_p(\underset{\sim}{s}_1)\}, \tag{2.10}$$

where the expression in braces represents the i-th row of the (n x p) matrix F.

Then,

$$\underset{\sim}{Z} = F\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}, \tag{2.11}$$

where $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$, and $\operatorname{var}(\underset{\sim}{\varepsilon}) = I\sigma_1^2$. This is a linear model with unknown parameters $(\underset{\sim}{\beta}, \sigma_1^2)$.

The primary goal is not to estimate these parameters but rather to predict $Z(\underset{\sim}{s}_0)$; $\underset{\sim}{s}_0\in D$. It is easy to see that the best (i.e., minimum mean-squared error) linear unbiased predictor is

$$\hat{Z}(\underset{\sim}{s}_0) = \underset{\sim}{f}(\underset{\sim}{s}_0)'\hat{\underset{\sim}{\beta}}, \tag{2.12}$$

where $\underset{\sim}{f}(\underset{\sim}{s}_0) = (f_1(\underset{\sim}{s}_0),\ldots,f_p(\underset{\sim}{s}_0))'$, and

$$\hat{\underset{\sim}{\beta}} = (F'F)^{-1}F'\underset{\sim}{Z}, \tag{2.13}$$

the best linear unbiased estimator of $\underset{\sim}{\beta}$. This predictor has mean squared error,

$$\operatorname{mse}(\hat{Z}(\underset{\sim}{s}_0)) = E(\hat{Z}(\underset{\sim}{s}_0) - Z(\underset{\sim}{s}_0))^2$$
$$= (\underset{\sim}{f}(\underset{\sim}{s}_0)'(F'F)^{-1}\underset{\sim}{f}(\underset{\sim}{s}_0) + 1)\sigma_1^2 \; . \tag{2.14}$$

2.2 Random field model

Recall that the trend surface model specifies that two data values, no matter how close they are to each other, are statistically independent. Most scientists working in geology, soil science, crop science, forestry, environmental science, etc., find this assumption contradicts their underlying notions of the mechanisms that resulted in the data they are studying. Put simply, in many scientific studies where data have a spatial label, it is felt that observations closer together are more dependent than those further apart. This is modeled as

$$Z(\underset{\sim}{s}) = \mu_2(\underset{\sim}{s}) + \delta_2(\underset{\sim}{s}); \quad \underset{\sim}{s}\in D, \tag{2.15}$$

where $\mu_2(\cdot)$ is the large-scale, deterministic, mean structure of the process, and $\delta_2(\cdot)$ is the small-scale, stochastic structure that models the spatial dependence referred to above. That is,

$$E(\delta_2(\underset{\sim}{s})) = 0 \qquad ; \; \underset{\sim}{s}\in D$$
$$\operatorname{cov}(\delta_2(\underset{\sim}{s}), \delta_2(\underset{\sim}{u})) \equiv C(\underset{\sim}{s},\underset{\sim}{u}) \; ; \; \underset{\sim}{s},\underset{\sim}{u}\in D. \tag{2.16}$$

The representation (2.15) is precisely the sort of approach time series analysts take; after detrending, they model the small-scale variation with stochastic processes that capture the temporal dependence (e.g., autoregressive, integrated, moving average processes).

In order to estimate the quantity and quality of the spatial dependence, some minimal assumptions of stationarity have to be made. A process $Y(\cdot)$ which satisfies

$$E(Y(\underset{\sim}{s})) = \mu \quad ; \; \underset{\sim}{s}\in D \tag{2.17}$$

$$\operatorname{var}(Y(\underset{\sim}{s}) - Y(\underset{\sim}{u})) = 2\gamma(\underset{\sim}{s} - \underset{\sim}{u}); \; \underset{\sim}{s},\underset{\sim}{u}\in D, \tag{2.18}$$

is said to be intrinsically stationary (over D). The quantity $2\gamma(\cdot)$ is a function of the vector difference $\underset{\sim}{s} - \underset{\sim}{u}$, and is called the variogram ($\gamma(\cdot)$ is called the semivariogram).

Intrinsically stationary processes are more general than second-order stationary processes, which assume

$$\operatorname{cov}(Y(\underset{\sim}{s}), Y(\underset{\sim}{u})) = C(\underset{\sim}{s} - \underset{\sim}{u}),$$

where $C(\cdot)$ is a function of the vector difference and is usually called the autocovariance function. It is the preferred tool of time series analysts, but for prediction purposes it is not as flexible as the variogram; see Cressie (1988) for more details and a summary of the properties of the variogram.

The geostatistical method (Matheron, 1963) of spatial prediction is very simple. Estimate the variogram with a nonparametric estimator $\hat{2\gamma}(\underset{\sim}{h})$, fit a model $2\gamma(\underset{\sim}{h};\underset{\sim}{\theta})$ whose form is known

apart from a few parameters $\underset{\sim}{\theta}$ (i.e., fit $2\gamma(\underset{\sim}{h};\underset{\sim}{\theta})$ to $2\hat{\gamma}(\underset{\sim}{h})$), and finally use the fitted variogram in the best linear unbiased predictor (i.e., the kriging predictor) as if it were known. However it is more the exception than the rule that $Z(\cdot)$ is intrinsically stationary; Cressie (1986, 1987a) investigates how to proceed when respectively $\mu_2(\cdot)$ is not constant and $\delta_2(\cdot)$ does not possess a variogram.

Let us assume for the moment that in (2.15) $\mu_2(\underset{\sim}{s}) \equiv \mu_2$, and $\delta_2(\cdot)$ is a zero mean intrinsically stationary stochastic process. Matheron (1963) has proposed the following variogram estimator

$$2\hat{\gamma}(\underset{\sim}{h}) \equiv (1/|N(\underset{\sim}{h})|)\sum_{N(\underset{\sim}{h})} (Z(\underset{\sim}{s}_i) - Z(\underset{\sim}{s}_j))^2, \quad (2.19)$$

where $N(\underset{\sim}{h}) \equiv \{(\underset{\sim}{s}_i,\underset{\sim}{s}_j): \underset{\sim}{s}_i - \underset{\sim}{s}_j = \underset{\sim}{h}\}$ and $|N(\underset{\sim}{h})|$ is the number of distinct ordered pairs in the set $N(\underset{\sim}{h})$. If the data are believed to be approximately Gaussian but with haphazard "contaminated" observations present, Cressie and Hawkins (1980) advocated the robust estimator,

$$2\bar{\gamma}(\underset{\sim}{h}) \equiv \{(1/|N(\underset{\sim}{h})|)\sum_{N(\underset{\sim}{h})} |Z(\underset{\sim}{s}_i) - Z(\underset{\sim}{s}_j)|^{1/2}\}^4 / \{0.457 + 0.494/|N(\underset{\sim}{h})|\}.$$

To either of these is fit a parametric model $2\gamma(\underset{\sim}{h};\underset{\sim}{\theta})$. Cressie (1985) proposes a weighted and a generalized least squares approach, while Kitanidis (1983) proposes a restricted maximum likelihood (REML) approach (i.e., estimation of $\underset{\sim}{\theta}$ is obtained from generalized increments of the data). Both methods have their strengths.

Assume that the estimation and fitting of the variogram has been carried out successfully, and that the variogram can be considered as known. Suppose it is desired to predict $Z(\underset{\sim}{s}_0)$ at some unsampled spatial location $\underset{\sim}{s}_0$ using a linear function of the data:

$$\hat{Z}(\underset{\sim}{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\underset{\sim}{s}_i), \quad (2.20)$$

which is unbiased (i.e., $E(\hat{Z}(\underset{\sim}{s}_0)) = E(Z(\underset{\sim}{s}_0))$) and minimizes the mean-squared error

$$E(Z(\underset{\sim}{s}_0) - \hat{Z}(\underset{\sim}{s}_0))^2.$$

Then the optimal λ's satisfy an (n+1)-dimensional linear equation that depends on $2\gamma(\underset{\sim}{s}_i - \underset{\sim}{s}_j)$; i,j = 0,1,...,n:

$$\Gamma \underset{\sim}{\lambda} = \underset{\sim}{\gamma}, \quad (2.21)$$

where $\underset{\sim}{\lambda} = (\lambda_1,\lambda_2,\ldots,\lambda_n,m)'$,

$\underset{\sim}{\gamma} = (\gamma(\underset{\sim}{s}_1 - \underset{\sim}{s}_0),\ldots,\gamma(\underset{\sim}{s}_n - \underset{\sim}{s}_0), 1)'$, and

$$\Gamma = \begin{cases} \gamma(\underset{\sim}{s}_i - \underset{\sim}{s}_j), & i = 1,\ldots,n,\ j = 1,\ldots,n \\ 1, & i = n+1,\ j = 1,\ldots,n \\ 0, & i = n+1,\ j = n+1; \end{cases}$$

Γ is a symmetric (n + 1)x(n + 1) matrix. The minimized mean-squared error is given by,

$$\sigma_k^2(\underset{\sim}{s}_0) = \sum_{i=1}^{n} \lambda_i \gamma(\underset{\sim}{s}_i - \underset{\sim}{s}_0) + m. \quad (2.22)$$

Matheron (1963) calls this spatial prediction method, (ordinary) kriging, after D. G. Krige (a South African mining engineer who in the 1950s became aware that predicted grade distributions were not a good indicator of actual grade distributions). In fact the optimal linear predictor (2.20) can be found earlier in the literature in works of Kolmogorov (1941) and Wiener (1949). The kriging predictor given by (2.21) assumes the process $Z(\cdot)$ contains no measurement error; Cressie (1987b) has details on how to predict noiseless versions of $Z(\cdot)$.

Several important design considerations become apparent from (2.21) and (2.20), viz. the optimal predictor and the minimized mean-squared error depend on the relative locations (2.2) of the data, the semivariogram values $\gamma(\cdot)$ and the sample size. These considerations will be picked up again in Section 3.

2.3 Comparison of trend surface model and random field model

Agterberg (1984) describes a small experiment to compare prediction based on a quadratic trend surface, prediction based on ordinary kriging, and prediction based on a quadratic trend surface plus kriging the residuals. It is not a priori true that the latter prediction method would perform the best since overfitting may occur. Agterberg took a contour map of the terrain at the top of the Arbuckle formation in central Kansas, within a 30 mile by 30 mile area. Then he randomly chose spatial locations from that area to provide the data. Finally spatial locations for prediction were randomly chosen and relative mean-squared errors:

$$\frac{\sum_{\underset{\sim}{s}_0} (\hat{Z}(\underset{\sim}{s}_0) - Z(\underset{\sim}{s}_0))^2}{\sum_{\underset{\sim}{s}_0} (\bar{Z} - Z(\underset{\sim}{s}_0))^2} \cdot 100\%$$

were computed. From the limited experiment, trend surface prediction (26.2%) was slightly worse than ordinary kriging (23.3%), and both were dominated by the combination of the two (19.1%).

Laslett et al. (1987) performed a very interesting spatial experiment to compare the performance of several two-dimensional spatial prediction methods. They note that rival methods of prediction are based on different models, so that a proper comparison of methods cannot be made analytically or on artifically generated data. They then propose to make the comparison on real data sets, viz. soil pH. Among the eleven predictors (which included quadratic trend surface and kriging) used, kriging and Laplacian smoothing splines performed the best. In their study, the presence of dependent small-scale variation was strongly suggested, justifying the use of geostatistical methods for prediction.

2.4 Nonstationary mean structure

In Section 2.3, there was mention of a combination (trend-surface-plus-kriging) prediction method. In fact Matheron (1969) has called this universal kriging. Recall the random field model (2.15). Let the large-scale variation be given by

$$\mu_2(\underset{\sim}{s}) = \beta_1 f_1(\underset{\sim}{s}) + \ldots + \beta_q f_q(\underset{\sim}{s}).$$

If the covariance structure $C(\underset{\sim}{s},\underset{\sim}{u})$ in (2.16) is known it is a simple matter to estimate $\underset{\sim}{\beta} \equiv (\beta_1,\ldots,\beta_q)'$ by generalized least squares, and predict $Z(\underset{\sim}{s}_0)$ using a slightly more general form of the kriging equations given in Section 2.2. Also, if $\underset{\sim}{\beta}$ is known then one can perform ordinary kriging on the residuals $Z(\underset{\sim}{s}) - \mu_2(\underset{\sim}{s})$. Unfortunately neither is usually known, and experimental variograms based on residuals

$Z(\underset{\sim}{s}) - \underset{\sim}{f}(\underset{\sim}{s})'\hat{\underset{\sim}{\beta}}$ are biased estimators of

$2\gamma(\underset{\sim}{h}) \equiv \mathrm{var}(\delta_2(\underset{\sim}{s}) - \delta_2(\underset{\sim}{s} + \underset{\sim}{h}))$ (assuming $\delta_2(\cdot)$ is intrinsically stationary). These issues are discussed in more detail in Cressie (1986), who proposes an approach to estimating the large-scale (spatial) variation that is based on median polish (Tukey, 1977). In two dimensions, this is simply a decomposition

$$\mu_2(\underset{\sim}{s}) = c + a_x + b_y \; ; \; \underset{\sim}{s} = (x,y), \qquad (2.23)$$

whose components are estimated by successive removal of medians from rows and columns. Call the result $\tilde{\mu}_2(\underset{\sim}{s})$ and define

$$\tilde{\delta}_2(\underset{\sim}{s}) \equiv Z(\underset{\sim}{s}) - \tilde{\mu}_2(\underset{\sim}{s}). \qquad (2.24)$$

Then carry out ordinary kriging on $\tilde{\delta}_2(\cdot)$; median-based residuals are less biased than mean-based residuals. Although easier to implement on regular grids, this technique of median polish kriging is also applicable to irregular spatial data (e.g., atmospheric pollution networks of sites), as is evidenced by the analysis in Cressie and Read (1986).

3. Spatial Design of Networks

The random field model (2.15) will be assumed; henceforth it will be written as

$$Z(\underset{\sim}{s}) = \mu(\underset{\sim}{s}) + \delta(\underset{\sim}{s}). \qquad (3.1)$$

To simplify matters, it will be assumed that one potential observation is available at any spatial location. (This is not as restrictive as it sounds since replicates can be averaged.) Thus the amount of replication at any site will not be considered a design parameter.

When optimal network design is added to the goal of optimal spatial prediction (see Section 2), the problem can be posed as one of finding design points $\underset{\sim}{s}_1,\ldots,\underset{\sim}{s}_n$ and weights $\lambda_1,\ldots,\lambda_n$ to minimize

$$E(\{\sum_{i=1}^{n} \lambda_i Z(\underset{\sim}{s}_i) - Z(\underset{\sim}{s}_0)\}^2 \,|\, Z(\underset{\sim}{s}_1),\ldots,Z(\underset{\sim}{s}_n)), \qquad (3.2)$$

the posterior mean-squared error, or more simply to minimize the mean-squared error

$$E(\{\sum_{i=1}^{n} \lambda_i Z(\underset{\sim}{s}_i) - Z(\underset{\sim}{s}_0)\}^2), \qquad (3.3)$$

which is easier to work with. Upon making intrinsic stationarity and unbiasedness assumptions, the optimal $\underset{\sim}{\lambda}$ is given by (2.21) and the minimized mean squared error by (2.22). To emphasize its dependence on n, we write (2.22) as

$$\sigma_k^2(\underset{\sim}{s}_0)^{(n)} = \sum_{i=1}^{n} \lambda_i \gamma(\underset{\sim}{s}_i - \underset{\sim}{s}_0) + m,$$

where $\underset{\sim}{\lambda} = \Gamma^{-1}\underset{\sim}{\gamma}$; see (2.21). This is now a function only of design points $\underset{\sim}{s}_1,\ldots,\underset{\sim}{s}_n$, and of n.

Suppose it is desired to add an (n+1)-th site. Where is the optimal location $\underset{\sim}{s}_{n+1}$? Define

$$S_n^2 \equiv \int_D \sigma_k^2(\underset{\sim}{s}_0)^{(n)} \, w(\underset{\sim}{s}_0;\, \sigma_k^2(\underset{\sim}{s}_0)^{(n)}) d\underset{\sim}{s}_0 \qquad (3.4)$$

to be a measure of the "average" prediction error. For example, suppose

$$w(\underset{\sim}{s}_0;\, \sigma_k^2(\underset{\sim}{s}_0)^{(n)}) = I(\underset{\sim}{s}_0 \in B).I(\sigma_k^2(\underset{\sim}{s}_0)^{(n)} > L^2), \qquad (3.5)$$

where I(A) is an indicator function that is equal to one if A is satisfied and is equal to zero otherwise. The "weight" function w focuses attention on the prediction error in a subregion B of D, and on only those prediction errors beyond a certain prespecified threshold L.

Adding another site at $\underset{\sim}{s}_{n+1}$ will decrease the average prediction error, provided new optimal $\lambda_1,\ldots,\lambda_n,\lambda_{n+1},m$ are chosen. I shall emphasize the addition of this new point by notating the resulting average prediction error in (3.4) as $S_{n+1}^2(\underset{\sim}{s}_{n+1})$. Then minimize this quantity over the location $\underset{\sim}{s}_{n+1} \in D$.

In fact, for reasons such as the logistics of setting up and then monitoring, the suitability of local conditions etc., there would probably be just a small number of potential sites.

Notice that

$$\frac{S_n^2(\underset{\sim}{s}_n)}{\sigma^2} = \frac{S_n^2(\underset{\sim}{s}_n)}{S_{n-1}^2(\underset{\sim}{s}_{n-1})} \, \frac{S_{n-1}^2(\underset{\sim}{s}_{n-1})}{S_{n-2}^2(\underset{\sim}{s}_{n-2})} \cdots \frac{S_1^2(\underset{\sim}{s}_1)}{\sigma^2}. \qquad (3.6)$$

Therefore, by starting on the far right of (3.6) and optimizing successively over each factor, a design could be found by adding one point at a time; the (i+1)-th site is placed in an optimal location having already fixed the i previous sites. Is this iterated design optimal?

The answer to the above question is, "no".

A very simple counter example is given by Cressie (1978), for $Z(\cdot)$ a process of independent increments on the real line. However, if one defines the design measure to be the limiting ($n \to \infty$) density of design points over the design space, then does the asymptotic iterated design measure equal the asymptotic optimal design measure? This is an open problem, but the answer is probably, "yes"; see Wu and Wynn (1978).

Although the iterated design may produce design points

$$V \equiv \{ \underset{\sim}{s}_1, \ldots, \underset{\sim}{s}_n \} \tag{3.7}$$

that are quite different from optimal design points when n is finite, it still offers the possibility of providing a suitable yardstick for other, more computer-intensive, algorithms. One such is the annealing algorithm, adapted by Sacks and Schiller (1987) for the spatial design problem. Suppose n is fixed; the problem is to find the best V. Again modify the notation of (3.4) and write S_n^2 as $S^2(V)$ to show its dependence on the design points.

The annealing algorithm assumes a starting value $V^{(0)}$. At the j-th stage, the current value is

$$V^{(j)} \equiv \{ \underset{\sim}{s}_1^{(j)}, \ldots, \underset{\sim}{s}_n^{(j)} \} \tag{3.8}$$

The algorithm proceeds as follows:

- Choose $V^{(j+1)}$ in some random or systematic manner (e.g., single point replacement)
- Replace $V^{(j)}$ by $V^{(j+1)}$ with probability π_j, where

$$\pi_j = \begin{cases} 1 & ; \text{ if } S^2(V^{(j+1)}) - S^2(V^{(j)}) < 0 \\ \exp\{-\log j (S^2(V^{(j+1)}) - S^2(V^{(j)})/\gamma\} & ; \text{ else .} \end{cases}$$

Notice that the algorithm always demands a change provided there continues to be an improvement in the objective function S^2. For suitable choice of γ in π_j, it will always avoid getting stuck at local minima by allowing a positive (ever-decreasing) probability of changing even when no improvement in the objective function is achieved. Geman and Geman (1984) is a good source for the properties of this stochastic relaxation technique.

But the question remains. At what stage should the algorithm be stopped? I suggest that the iterated optimal S_n^2 obtained from (3.6) could be used as a yardstick below which the quantity $S^2(V^{(j)})$ must fall. Other two at-a-time, etc. iterated designs that have rather straightforward computational solutions, could also be used as yardsticks.

Finally, the question of how many sites should be included in the network, needs to be addressed. As always, there are conflicting forces at work here; too many sites and the monitoring program will bankrupt itself, too few sites and the information is not adequate enough to allow definitive conclusions. One way to approach this is to take a predictor and calculate its average mean-squared error (3.4) under a special case of the model (3.1) where the data $Z(\underset{\sim}{s}_1), \ldots, Z(\underset{\sim}{s}_n)$ are independent, and under the model (3.1) where the data $Z(\underset{\sim}{s}_1), \ldots, Z(\underset{\sim}{s}_n)$ are dependent through the error random field $\delta(\cdot)$. Call the former $S^2_{n,\varepsilon}$ and the latter $S^2_{n,\delta}$. Then the effective number of observations neff is obtained by solving

$$S^2_{\text{neff},\delta} = S^2_{n,\varepsilon} . \tag{3.9}$$

Thus the network designer could think about having "neff" independent observations and make decisions about precision and cost as if the data were independent. When an "neff" has been decided, it can be converted back to the proper "currency" using (3.9).

A simple example would be to use the predictor $\bar{Z}$ to construct (3.9). Suppose there is an $\underset{\sim}{s}_0$ of particular interest. Then the above considerations lead to

$$\text{neff} = n\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}(Z(\underset{\sim}{s}_i), Z(\underset{\sim}{s}_j))/n - 2 \sum_{i=1}^{n} \text{cov}(Z(\underset{\sim}{s}_i), Z(\underset{\sim}{s}_0))\}.$$

It would be dangerous to demand too much from this rather heuristic approach, since "neff" very much depends on the criterion chosen in (3.9). In other words, an "neff" constructed according to one criterion but used for a different purpose in another criterion, may yield nonsensical results.

Acknowledgement

This research was supported by the National Science Foundation under grant number DMS-8703083.

References

Agterberg, F. P. (1984). Trend surface analysis, in Spatial Statistics and Models, eds. G. L. Gaile and C. J. Willmott. Reidel, Dordrecht, 147-171.

Box, G. E. P., and Jenkins, G. M. (1970). Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco.

Cressie, N. (1978). Estimation of the integral of a stochastic process. Bulletin of the Australian Mathematical Society, 18, 83-93.

Cressie, N. (1985). Fitting variogram models by weighted least squares. Journal of the International Association for Mathematical Geology, 17, 563-586.

Cressie, N. (1986). Kriging nonstationary data. Journal of the American Statistical Association, 81, 625-634.

Cressie, N. (1987a). A nonparametric view of generalized covariances for kriging. Mathematical Geology, 19, 425-449.

Cressie, N. (1987b). Spatial prediction and ordinary kriging. Proceedings of the Mathematical Geologists of the United States Meeting (Redwood City, CA, April 1987), forthcoming.

Cressie, N. (1988). Variogram, in Encyclopedia of Statistical Sciences, eds. S. Kotz and N. L. Johnson. Wiley, New York, forthcoming.

Cressie, N., and Hawkins, D. M. (1980). Robust estimation of the variogram: I. Journal of the International Association for Mathematical Geology, 12, 115-125.

Cressie, N., and Read, T. R. C. (1986). Spatial data analysis of regional counts. Statistical Laboratory Preprint No. 86-46, Iowa State University, Ames.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrological applications. Water Resources Research, 19, 909-921.

Kolmogorov, A. N. (1941). Interpolation and extrapolation of stationary random sequences. Izvestiia Akademii Nauk SSR, Seriia Matematicheskaia, 5, no. 3.

Laslett, G. M., McBratney, A. B., Pahl, P. J., and Hutchinson, M. F. (1987). Comparison of several spatial prediction methods for soil pH. Journal of Soil Science, 38, 325-341.

Matheron, G. (1963). Principles of geostatistics. Economic Geology, 58, 1246-1266.

Matheron, G. (1960). Le Krigeage Universel, Cahiers du Centre de Morphologie Mathématique, no. 1, Fontainebleau, France.

Peters, N. E., and Bonelli, J. E. (1982). Chemical Composition of Bulk Precipitation in the North-Central and Northeastern United States, December 1980 through February 1981. U. S. Geological Survey Circular 874.

Sacks, J., and Schiller, S. (1987). Spatial designs, in Proceedings of the Fourth Purdue Symposium on Decision Theory and Related Topics, eds. S. S. Gupta and J. O. Berger. Springer, New York, forthcoming.

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley, Reading.

Wiener, N. (1949). Extrapolation, Interpolation and Smoothing of Stationary Time Series. MIT Press, Cambridge.

Wu, C-F. J., and Wynn, H. P. (1978). The convergence of general step-length algorithms for regular optimum design criteria. Annals of Statistics, 6, 1273-1285.

SPATIAL AUTOCORRELATION: IMPLICATIONS FOR SAMPLING AND ESTIMATION

Evan J. Englund
U.S. Environmental Protection Agency, P.O. Box 93478, Las Vegas, NV 89193-3478

ABSTRACT

Spatial autocorrelation is a normal result of the physical and chemical processes which operate in the environment. All measurable environmental parameters will exhibit spatial autocorrelation at some scale. This not only causes technical problems in sampling and estimation, but leads to more fundamental problems in communication. Terms such as "chemical concentration", "representative sample", and "frequency distribution" which are commonly used and well understood in laboratory situations, can become essentially meaningless when applied to environmental measurements without an explicit statement of the spatial scale (support) being considered.

A simulated spatially autocorrelated distribution is used to illustrate the changes in concentration, frequency distribution, and sample quality associated with changes in support. Variograms computed from the simulated data illustrate the relationship between spatial variability and standard QA/QC. Practical suggestions are made for sampling and estimation of spatially autocorrelated sites.

INTRODUCTION

The purpose of environmental sampling programs, like other samplings, is to use the information obtained from the sample to make inferences about the larger population from which the sample is drawn. For example, an industrial waste lagoon may be sampled to determine whether its mean concentration exceeds an allowable maximum value for a particular chemical. In other cases, sampling may be done to determine whether the level of pollution has increased or decreased from previous levels at a site, or to identify the locations of polluted areas. A typical sequence of events in an environmental investigation might include the design of a sampling plan, collection of samples, laboratory chemical analysis, interpretation of data, and finally, a decision based on the interpretation.

Decisions made by various federal, state, and local agencies may require remedial, preventative, or punitive measures which may have substantial consequences in terms of human health as well as economics. It is obviously important that environmental investigations be conducted in such a manner as to ensure that decisions are based on the best possible information. It should be emphasized here that the information on which decisions are made is not the sample data itself, but the interpretation of the data; that is, the estimates of the characteristics of the larger population which are made from the data.

Unfortunately, the interpretation of field data is made difficult by the nature of the environment itself, which is not always amenable to the methods of sampling and data analysis designed for use under more controlled conditions such as laboratory experiments. Although the same sampling and data analysis terminology is used in both the laboratory and the field, the precise meanings of many terms are significantly different. The resulting ambiguities can lead to problems in both the design of sampling programs, and the interpretation of the results. This paper will explore the causes of these problems, attempt to clarify the terminology, and offer some practical suggestions for sampling design and data interpretation.

SPATIAL AUTOCORRELATION

Spatial autocorrelation is the basic cause of the problems to be discussed. All measureable environmental parameters exhibit spatial (and temporal) autocorrelation at some scale. This means that over some range of distances, measurements will tend to be more similar to measurements taken nearby than to measurements taken farther away. For practical purposes, stating that a phenomenon is spatially autocorrelated is equivalent to stating that it is not uniformly distributed, but autocorrelation is more easily quantified in statistical terms. The term spatial correlation is often used interchangeably with spatial autocorrelation, but the former implies that the measured values are correlated with their locations, as temperature, for example, is correlated with distance from the equator.

The physical and chemical processes that control the fate and transport of chemicals in the environment do not operate at random, although most events include what may be considered random processes, in the sense that they are too complex to be predicted in detail. Random and deterministic processes may operate simultaneously at several scales to produce the phenomenon being measured. The measurement of precipitation (with or without acid), provides a good illustration. Regional weather patterns produce the conditions necessary for rainfall over a large area. Within the area, individual clouds or clusters of clouds form, apparently at random locations, due to local fluctuations in temperature and humidity. If one collected pairs of rain-guage measurements taken at various spacings, one would expect to observe strongly correlated readings at separations much smaller than the size of the average cloud, weaker correlation at regional-scale separations, and essentially zero correlation for separations larger than the size of the region. Although contiguous rain guages would be expected to show the greatest correlation, it would not be perfect, due to the fact that the two guages may not receive identical amounts of rain (spatial variability), and even if they did, the readings may not be identical (measurement error).

After a major rainfall, for example, we might be asked to estimate the average precipitation over a watershed from a given set of rain guage readings, in order to estimate the watershed's contribution to a flood crest. Alternately, we might want to know how many guages we would need in order to predict flooding with a specified degree of accuracy. The methodology commonly known as geostatistics (Matheron, 1963; Journel and Huijbregts, 1978) was developed in the mining industry to deal with questions of local ore grade estimation, and has since been shown to be

generally applicable to most situations where spatial autocorrelation is present.

A basic assumption in geostatistical analysis is that the spatial autocorrelation exhibited in a set of measurements can be represented by an underlying autocorrelation function which is valid over the region of interest. For many environmental phenomena, this assumption can be intuitively related to the controlling processes.

The variogram is one commonly used method for quantifying spatial autocorrelation (Fig.1). Experimental variograms are computed from sample data by examining all possible data pairs, grouping them by distance classes, and computing a variance for each distance class using the standard paired-sample formula. A theoretical model is then fitted to the experimental data. A variogram contains exactly the same information as a plot of correlation coeficients for the same distance classes (one can be transformed into the other by inverting the plot and rescaling the y-axis).

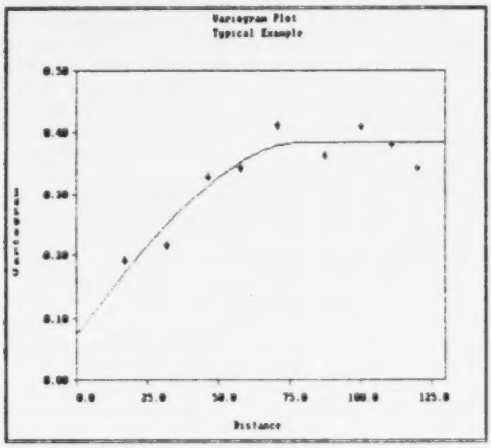


Fig. 1. Typical variogram plot and fitted model. Variances are computed from paired sample differences for pairs in successive distance classes and plotted against distance. The fitted model exhibits commonly observed features: a random component or "nugget" at the y-axis intercept, and an increase in variance with distance up to a maximum "range" of autocorrelation.

SAMPLE SUPPORT

When spatial autocorrelation is present, the physical dimensions of the sample become important considerations. The term 'support', defined as the "size, shape, and orientation of the physical sample taken at a sample point" (Starks,1986), is used to avoid confusion with the statistical size (number of observations) of a sample. The support of a soil core, for example , would be its diameter and length. In the case of a rain guage measurement, the support would include the diameter of the orifice, and the time of accumulation.

When the sample support changes, the statistical properties of the sample set such as the variance of samples and the sampling variance also change. These changes make sample support a critical element in the design of a sampling program. The concept of support applies equally to the stage of data interpretation. The support on which decisions will be made is rarely the same as that of the samples. The choice of decision support can significantly affect the outcome of an analysis, and should be considered before a sampling program is undertaken. The idea of support and its ramifications will be developed further in the following sections.

CHEMICAL CONCENTRATION

The term 'chemical concentration' is meaningless in the absence of a specified support. Atoms and molecules represent the smallest scale at which elements and compounds can be said to exist. If samples are taken at this scale, the true concentration of any substance within the sample will be a discrete binary phenomenon - the concentration will be either 100% or 0%. Any larger sample is made up of a mixture of discrete components, and its true concentration will be the sum of the weight or volume of the analyte divided by the total weight or volume of the sample; i.e., the average over the sample support. Usually the entire volume of a sample is not measured directly, but the measured concentration of a subsample is used to estimate the the mean concentration of the original sample. Great care is taken during the preparation of a sample to ensure uniform mixing, and if necessary, to reduce the particle size of the material so that any subsample used for analysis will have a true mean concentration very close to the true mean of the entire sample. If the mixing is effective, neither the size of a subsample nor its location should have a significant effect on the outcome of the analysis...for practical purposes, one subsample is as good as any other.

In the field, where the area being investigated cannot be uniformly mixed, the situation is quite different. Layers, crystals, clumps, or other high concentrations of a substance often occur such that if a given sample had been taken at a slightly different location (sometimes only a fraction of the sample size away), a significantly different true sample concentration would be found. The classic example of this was observed in placer gold deposits, where the presence or absence of a single, miniscule gold nugget in a sample would make the difference between an assay indicating high-grade ore, and one indicating waste. This led to the interesting term 'nugget effect' often being applied to the y-axis discontinuity in variogram models.

Like changes in location, changes in support also result in changes in concentration. The true value of a point sample is 0 or 100%; the true value of any larger volume centered on the point is the mean concentration over the volume. If the dimensions of a sample are increased or decreased; if the shape of the sample is changed, say from a sphere to a cube; or if the orientation of a non-spherical shape is changed; the sample will contain a different set of molecules, and probably a different true concentration. For any point in space which represents a potential sample 'location', an infinite number of possible sample supports exist centered on the point, and an infinite number of possible true concentrations which can be said to be the concentration at that location. Obviously, we must conclude that any reported measurement of chemical concentration in the environment is essentially meaning-

less unless the support is also reported. Likewise, a statement such as "remedial action will be taken if the concentration of cadmium in soil exceeds 500 ppm" is also meaningless unless the support is specified.

A SIMULATED EXAMPLE

An example based on a simple computer simulation serves to illustrate the support problem. A blank computer screen is 'polluted' with a 'realistic-looking' pattern of pixels (Fig. 2). The algorithm which was used to generate the pattern first selected 25 points at random in the central part of the screen. Each of these points was used as the center of a cluster of up to 2000 points scattered around the center at random angles and approximately normally distributed distances (sum of three uniformly distributed random values). Blank pixels were initialized with a value of zero, and incremented by one each time they were hit by the point generator. A color terminal can be used for a more effective display than Fig. 2, because pixel values greater than one can be represented by various color codes. The details of the algorithm are not crucial. Any algorithm which conditions the outcome of a random process on a prior outcome of the same or another random process, can be used to generate spatially autocorrelated patterns. Variations on the drunkard's walk (e.g., two steps forward, three steps back) are effective.

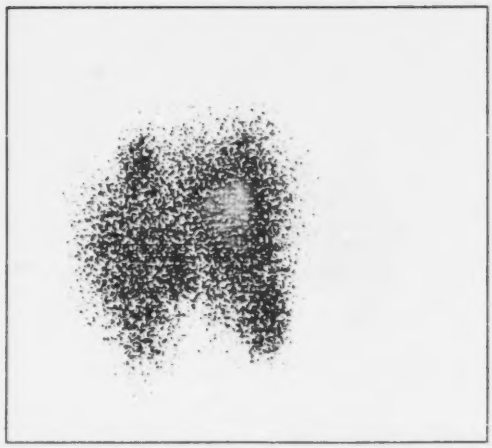


Fig. 2. A simulated, spatially autocorrelated distribution of "pollutant".

When the pattern has been generated, we have an area subdivided into pixel-sized units, for each of which we know the exact pollution concentration. The pixel scale of support is considered to be smaller than any support size we would be interested in, but since we have exact knowledge at that scale, we can now combine pixels into any larger support areas we choose, and compute the exact average pollution concentrations over these areas.

Figs. 3-7 illustrate the results of this kind of change in support. In Fig. 3, the screen area was divided into support blocks of 5x5 pixels, and the mean pixel value was computed for each block. The block means were grouped into class intervals and represented as shaded patterns on the map. The histogram scaled by area, and univariate statistics, are also shown for the set of non-zero value blocks. Figs. 4-7 repeat this process at supports of 10x10, 20x20, 40x40, and 80x80 pixels, respectively.

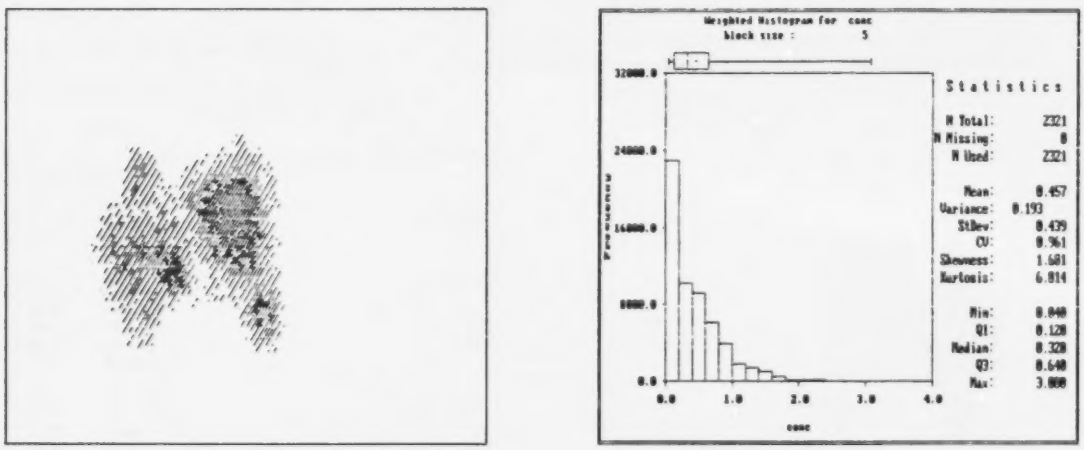


Fig. 3. Map and histogram of true "concentrations" of simulated pollutant averaged over 5x5 pixel blocks. Darker shades on map represent higher values: Blank (<0.4 units per pixel); Light shade (0.4 to 0.8); Medium shade (0.8 to 1.2); Dark shade (>1.2).

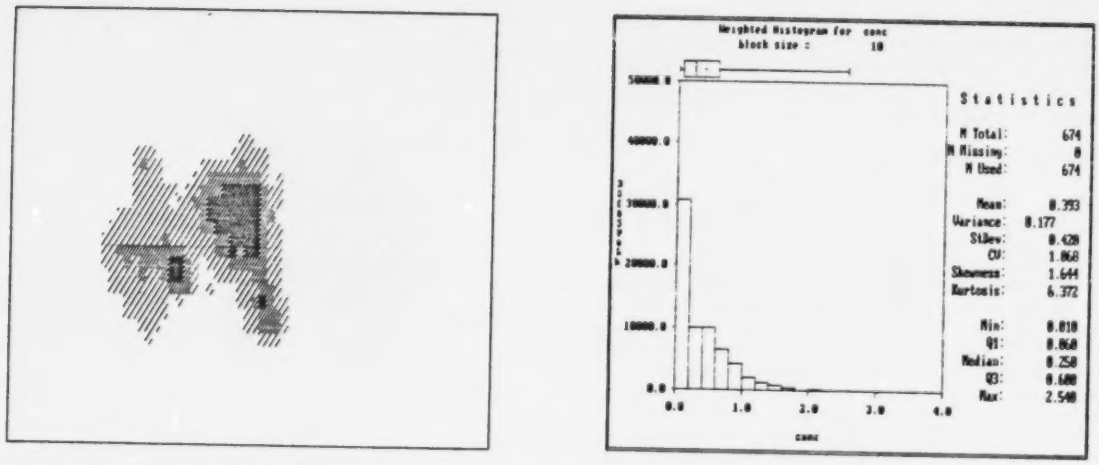


Fig. 4. Map and histogram of true "concentrations" of simulated pollutant averaged over 10x10 pixel blocks. Shades represent the same values as in Fig. 3.

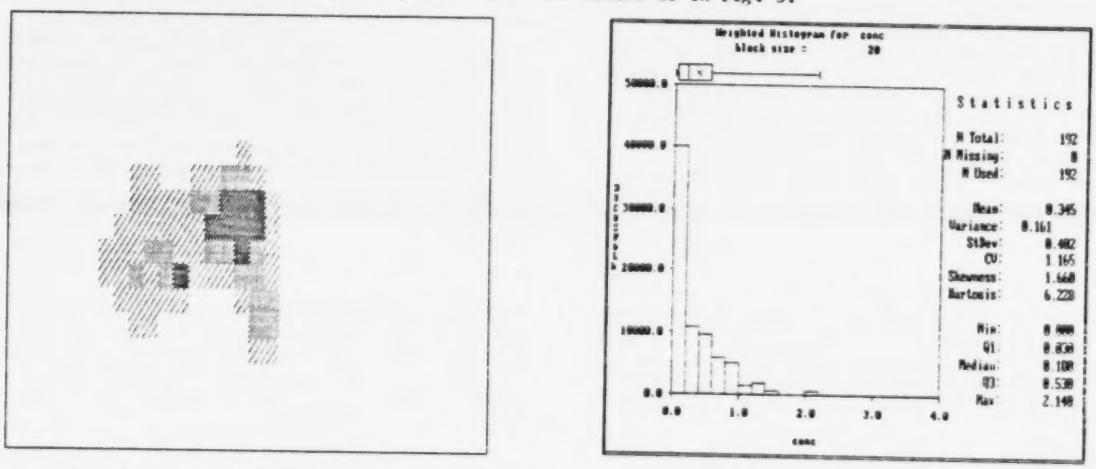


Fig. 5. Map and histogram of true "concentrations" of simulated pollutant averaged over 20x20 pixel blocks. Shades represent the same values as in Fig. 3.

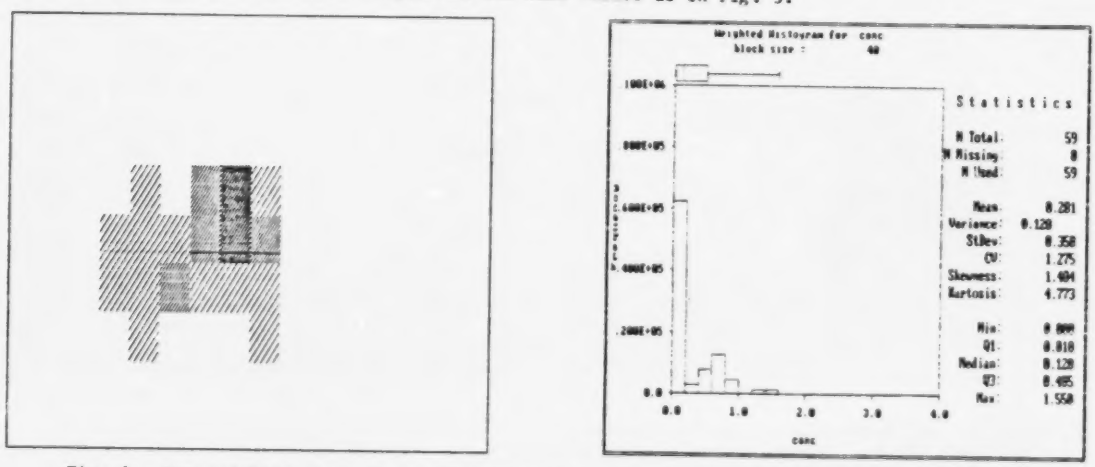


Fig. 6. Map and histogram of true "concentrations" of simulated pollutant averaged over 40x40 pixel blocks. Shades represent the same values as in Fig. 3.

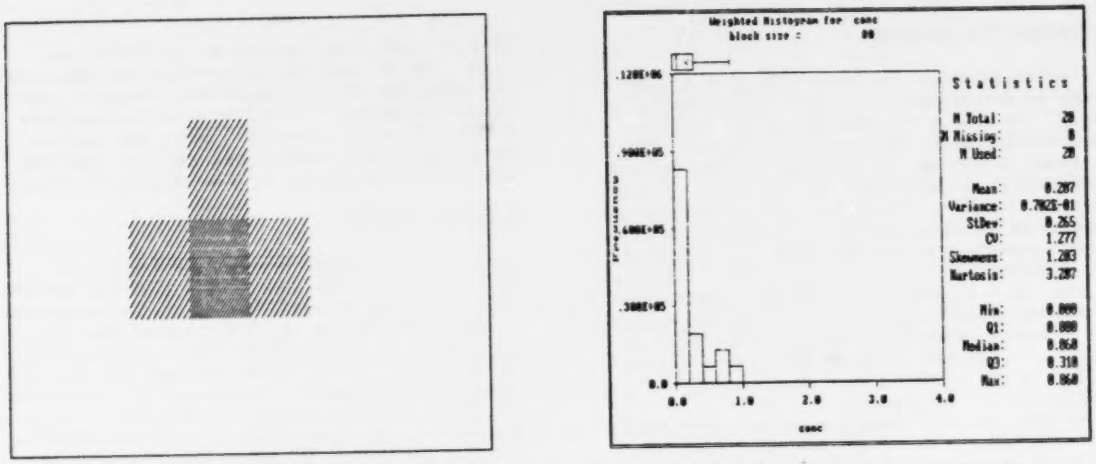


Fig. 7. Map and histogram of true "concentrations" of simulated pollutant averaged over 80x8 pixel blocks. Shades represent the same values as in Fig. 3.

The most significant thing to remember when examining these results is that these are not estimates of the underlying distribution. Each of these is the true distribution for the specified support and location. Additional, equally true distributions can be generated by shifting the grid origin, by filling in the missing block sizes (6x6, 7x7,...), by changing the shape (4x6, 8x12,...), etc. Continuing our analogy between the simulation and a polluted site, we must ask ourselves, as site investigators, which truth do we want to know? Or perhaps more realistically, which truth do we need to know, and what do we have to do to know it with sufficient accuracy?

If we want to take some remedial action, say, by cleaning up those areas in excess of an action level of 1.2, we must first specify the remediation support. If we specify 5x5 pixel blocks as in Fig. 3, our goal would be to identify as polluted, and clean up, all of the darkest blocks. If, on the other hand, we specify the 80x80 pixel blocks, we would find that there is nothing to clean up at all. If a support is not specified along with an action level in the applicable regulations, the site investigator is faced with the problem of establishing one, either explicitly or implicitly. Along with the flexibility comes the cost of ambiguity: long meetings spent arguing over semantics; or perhaps extensive litigation.

FREQUENCY DISTRIBUTIONS

As is evident from the maps and histograms in Figs. 3-7 when the support changes, the frequency distribution also changes. Within the domain defined by the screen boundaries, as the support dimension increases, the variance of the distribution decreases. This is particularily important in the context of regulation and remediation, because, as was suggested above, increasing the support dimension can often reduce (or eliminate) the amount of material above an action level. It is interesting to note that support is frequently specified in the time dimension for air quality monitoring and regulation; it is common to refer to 1-hour averages, 24-hour averages, etc., recognizing that these would give quite different results if compared against a single concentration limit.

From a practical standpoint, as mentioned above, specifying the support associated with an action level will reduce ambiguity in site assessment and enforcement. At a more fundamental level, support is a significant factor in the establishment of action levels. Environmental actions are ultimately decisions about risk abatement, and the risks associated with chemical substances are in turn functions of toxicity and exposure, both of which involve questions of support. If two one-hectare lots each have mean concentrations of 1000 ppm lead in the upper ten centimeters of soil, are the risks associated with them significantly different if the distribution of one-square-meter supports for one lot is normal with a standard deviation of 10, while the other is log-normal with a standard deviation of 800? If remedial action were practical only at the hectare support, should the action decision be based on the mean concentration over the hectare, on the frequency distribution of some smaller support within the hectare, or on some combination?

An observation worth noting about the frequency distributions in Figs. 3-7 is that although the variance of the distributions decreases as the size of support increases, the shape of the histogram seems to remain relatively constant. If the spatial distribution were random rather than autocorrelated, the variance would be expected to decrease faster as the support increased, and the shape of the histogram would be expected to converge more rapidly toward the normal. The persistance of histogram shape in spatially autocorrelated fields suggests that affine correction of variance (Journel and Huijbregts, 1978, p. 471) may be a useful tool for estimating the frequency distribution of one support, given the frequncy distribution of another.

REPRESENTATIVE SAMPLES

'Representative' is one of the most misused terms in environmental sampling (Splitstone, 1987). In this case, the effects of spatial autocorrelation in the field combine with inconsistency in the use of the term 'sample' itself, resulting in a great deal of confusion. A statistical sample of n observations is sometimes said to be representative if the mean, and or other parameters of its frequency distribution are 'similar' to the population from which it was drawn. In a somewhat more abstract sense, the distribution might be called representative if it was properly drawn (that is, if each member of the population had an equal chance of being included in the sample), even though it is not similar to the population.

An environmental sample such as a soil core, collected in the field and sent to a lab for analysis, is not a 'sample' in the commonly used statistical sense. It might be more appropriately called a member of a population which has been included in a sample of that population. A set of n such environmental samples, or observations, make up a statistical sample which can be representative of the population of possible observations of the same support within the sampled domain. It can not, however, be directly representative of any other support population, except for the population mean, which is constant within the domain regardless of support. In the context of ordinary random statistics, it is therefore difficult to ascribe any representativeness to an individual environmental sample other than being a member of a representative set at that support.

In the case of environmental field investigations, however, the whole point of a sampling plan is often to obtain samples (observations) that are representative in a different sense, namely, representative of their local areas. This spatial kind of representativeness is also hard to define. Intuitively, what we are really after is a measurement at a location that is close to the local mean concentration, but as we have seen, the local mean is a function of the support. Thus, until we define the support which we want our observation to represent, we can't say anything about how good the observation is.

Given a set of n observations within a domain, a geometric support neighborhood can be defined around each observation as the set of all points closer to it than to any other observation (i.e., a Voroni polygon). Voroni polygons are an approximation of the neighborhoods with which the observations are most correlated, and give an idea of the spatial resolution of the set of observations.

APPLICATIONS

In most of the above discussion, we have dealt with the uncertainties and ambiguities involved in taking and measuring samples of the physical environment, and the inherent limitations of relating measured values back to some 'real' characteristic of the environment. In spite of these problems, most of which have not been adequately dealt with theoretically, the world marches on. Site investigators are still faced with the necessity of collecting samples in the field, and using them to make interpretations and decisions. The remainder of this paper will focus on practical approaches to sampling and data analysis which will help reduce ambiguity, and improve data quality and usefulness. We will continue to use the simulation as an illustrative example.

SAMPLING DESIGNS

One of the common problems facing a site investigator is the layout of a sampling pattern. Assume for the moment that you have a predefined domain that must be sampled and that the number of samples you can take has been fixed by budget constraints. Where should you take the samples? Most available guidance documents such as SW-846 recommend random sampling as the general solution. However, it has been shown (Olea, 1984, Yfantis, et al., 1987) that in the presence of spatial autocorrelation, sampling on a systematic grid will produce a more efficient sampling. If spatial autocorrelation is not present, the regular grid will be no better or worse than a random sample. Because the regular grid is a periodic sampling in space, it is obviously contraindicated when the presence of a spatial periodicity in the phenomenon is suspected at a scale near that of the proposed grid. Fortunately, this situation is not common, and the regular grid, because of its simplicity and effectiveness, can be used in most spatial sampling programs.

The efficiency of the regular grid is a result of the minimization of spatially clustered data which are duplicating each other's information. As a practical matter, small departures from regularity do not have major effects, so that field crews can make offsets from the grid to avoid obstacles without affecting the results of a study. This is particularly true when the autocorrelation function exhibits a large random component (nugget effect) relative to the autocorrelated component. The larger the relative nugget effect, the greater the tolerance area around the regular grid nodes.

SAMPLE SUPPORT AND QA/QC

To illustrate the potential impact of sample support on the quality of an investigation, we will sample our simulation at two different supports, and use the results for interpolation of concentrations. The sample locations are on the regular grid shown in Fig. 8. The first sampling is done at the single pixel support. Each sample is assigned the true value of the pixel at that location; we are assuming no sampling or analytical error. The second sampling is done on the same grid at a 2x2 pixel support. In this case the true mean value of the 4 pixels in the sample is used.

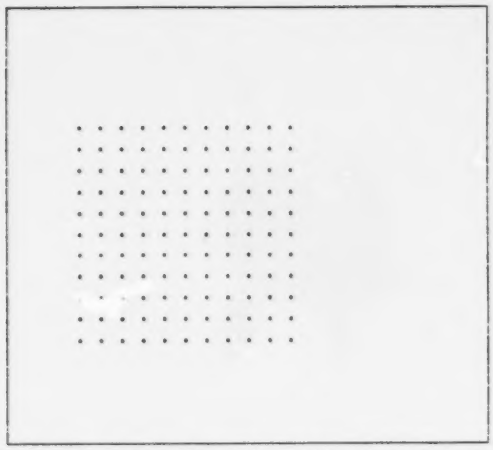


Fig. 8. Location of samples taken from the simulated distribution. True values of single pixels and 2x2 pixel blocks were taken at each point.

Variograms computed from the two sample sets are shown in Fig. 9. Note the dramatic decreases in variance at all distances with the larger support. When considering quality control, we take particular note of the projected y-axis intercept or nugget, which has dropped from a value of 0.5 to less than 0.1. The nugget value provides an estimate of the total variance associated with taking adjacent, or co-located samples,and includes small-scale spatial variability as well as sampling, preparation, and analytical errors. In the present example, we introduced no measurement errors, so all of the nugget value is the result of spatial variability, and the change in support has a major effect. If, at the other extreme, the nugget were entirely due to sampling and analytical errors, it would not be reduced at all by changing support.

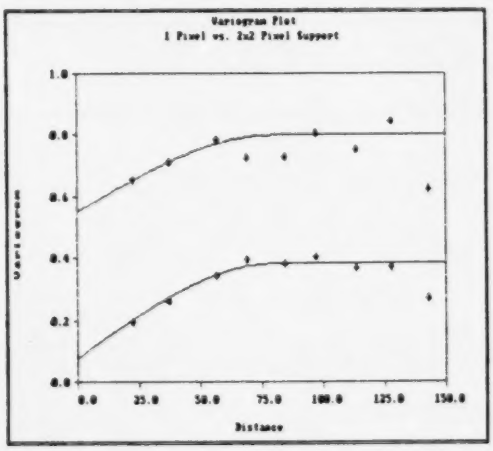


Fig. 9. Variograms computed from simulated samples on single pixel (upper curve) and 2x2 pixel (lower curve) support. Points are experimental values; solid lines are subjectively fitted models.

The approach one takes to improving data quality, therefore, should ideally be dictated by a variance components analysis of the variogram nugget value. The largest variance component should get the most attention. If spatial variance dominates, increase the support by taking larger samples or perhaps by compositing small clusters of samples. If measurement error dominates, look for a more accurate method. When either spatial or measurement variance is very dominant and cannot practically be reduced further, it may be possible to achieve a significant cost saving without seriously affecting overall data quality by going to cheaper methods which moderately increase the lower variance component. For example, if the maximum feasible support still results in a spatial variance component ten times greater than the measurement variance, the measurement variance could be allowed to increase by a factor of five with a resultant increase in the total standard error of only 17%.

The question of quality assurance goes beyond quality control of sample data, to include the adequacy of the program as a whole to achieve the desired objectives. In the case of a site investigation, we are particularily interested in the quality of the spatial estimate made from the sample data, and we must therefore look at sample quantity and location, and the estimation or interpolation procedures, as well as sample data quality. All of these factors can contribute to the quality of the end result, and as always, the most cost effective use of resources is to work at reducing the largest error component.

With our simulated example, we can compare the effects of the sample quality on the overall quality of the data interpretation by interpolating concentrations over the screen area from each of the two sample sets. Using the variogram models from Fig. 9, kriged estimates were computed for 10x10 pixel blocks. The results are shown in Fig. 10, with the blocks shaded according to the same classification used in Fig. 2. Comparison with the true values shown in Fig. 3 illustrates the overall superiority of the 2x2 sample support in defining the general pattern of block concentrations.

Plots of the true vs. estimated concentrations for the two interpolations are shown in Fig. 11, and histograms of the estimation errors in Fig. 12. Note that even though the 2x2 pixel samples provide better estimates, there is still relatively high error which may lead to a large number of false positives and false negatives at most concentration action levels. Further reduction of estimation error would require more samples, larger support, or both.

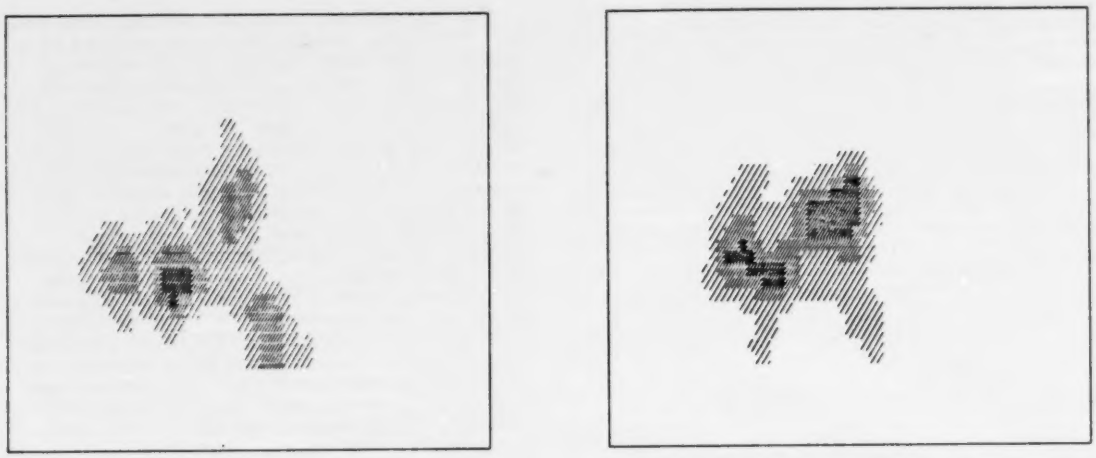


Fig. 10. Maps of kriged estimates for 10x10 pixel blocks using single pixel (left) and 2x2 (right) samples. Compare with map of true block values in Fig. 4.

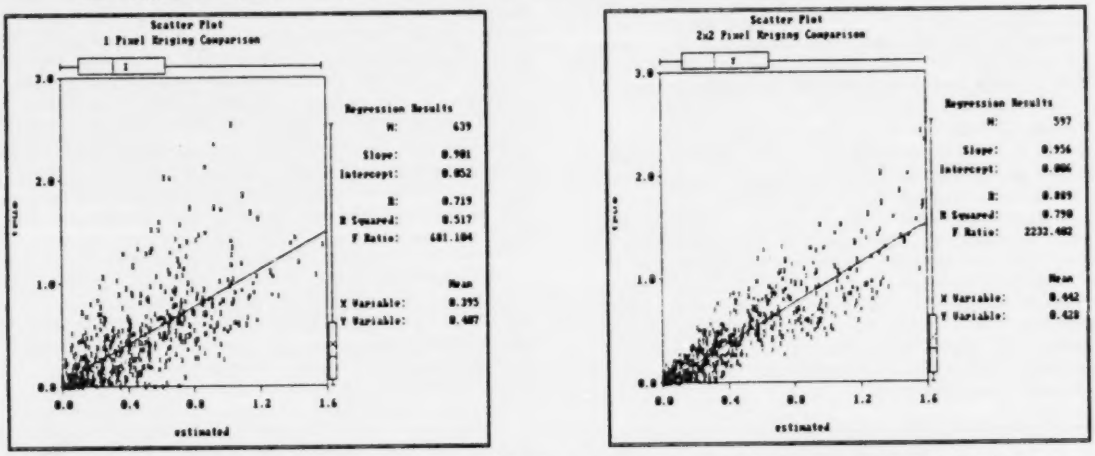


Fig. 11. Plots of estimated vs. true block values and associated regression statistics for the two kriged estimates in Fig. 10. The left plot is from the single pixel case, the right, from the 2x2 pixel case.

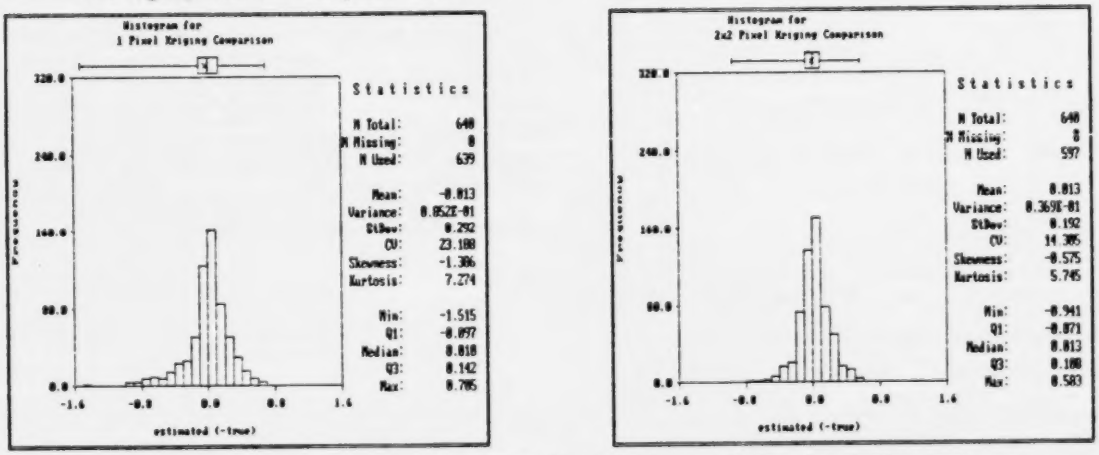


Fig. 12. Histograms of estimation errors (estimated - true) for the single pixel case (left), and the 2x2 pixel case (right).

ISOPLETH MAPS

Isopleth, or contour, maps are a commonly used and effective method for displaying the spatial distribution of chemical concentration. The use of isopleths, however re-introduces some of the ambiguity discussed above regarding support. If we examine the kriged maps in Fig. 10, we can easily visualize the process of generating a contour line by smoothing the stepwise boundary between two shading patterns. This is essentially the same result we would get if we assigned the estimated block average values to the block midpoints and ran the resulting grid through a typical contouring algorithm. It is also comparable to the limit we would approach (very expensively) with kriging as the grid dimension approaches zero. The problem arises because a contour represents the intersections of a plane parallel to the x-y axis and a continuous surface. The existence of one contour line implies the existence of all possible contour lines, and demonstrates that we have in effect estimated individual concentration values at every point. But we showed earlier that point concentrations are binary and discontinuous, and that meaningful statements about concentration require specifying a support. Which support?

Intuitively, we might like to think of the isopleth surface as representing the result of a moving average based on a specified support window. While this sounds good, it doesen't work in practice. When we specified a 10x10 block grid, it was easy to compute the true block means and compare them to the estimated block means. However, when we try to compute a true moving average isopleth based on a 10x10 support window, we find peaks, holes, and irregularities in the isopleth lines occurring at much smaller scales than the 10x10 support. Such a situation is self-contradictory, and it is not obvious how to define a true isopleth, compare it with an estimated one, and determine the goodness of the estimate.

In spite of the difficulty of defining a 'real' isopleth, a case can be made for using isopleths drawn from kriged block estimates as remediation boundaries. Given that kriging is in some sense an optimal estimator, and the goal is to remediate all of the area above an action level, the best approach should be to krige every point in the sampled domain, and remediate all of the points estimated to be above the action level. Contouring kriged block values provides a good approximation which is computationally feasable. The point to remember when doing this is that the nice, smooth contours may be primarily a regression effect, and that the underlying reality may in fact be very erratic. If we were to clean up an area inside an isopleth boundary, and then take check samples immediately around the boundary, we would expect to see a large number (approximately 30 to 50%, depending on the distribution) of samples exceed the action level, even if the boundary is well estimated. A more practical approach would be to do the check sampling along the boundary before the clean-up, and use the additional data to better define the boundary.

CONCLUSIONS

Chemical compounds in the environment can be spatially autocorrelated at scales ranging from the molecular to global. Autocorrelation is practically significant when it occurs at scales relevant to the problem, that is, between the dimensions of a sample and the dimensions of the domain being investigated. Under such conditions the support, or physical size, shape, and orientation of a sample, becomes a critical factor in the quality of an investigation. The use of QA/QC data for a variance components analysis of the nugget component of a variogram model can be very useful in selecting the most cost-effective approach to additional sampling.

Chemical concentration in the inhomogeneous environment must refer to a specific support to be meaningful. Specifying the support associated with an action level for enforcement or remediation removes a source of ambiguity and potential misunderstanding.

Spatial estimates of concentration such as kriged blocks, and the contour maps derived from them, are smoothed representations of reality, and represent a support more like the sampling grid size than the dimensions of the sample. Site investigators should understand, and be able to explain to the public, why many of the values at the sample support, taken outside an isopleth boundary, are expected to exceed the isopleth value.

NOTICE

The information in this document has been funded wholly by the U.S. Environmental Protection Agency. It has been subjected to Agency review and approved for publication.

REFERENCES

Journel, A. G., and Huijbregts, Ch. J., 1978. Mining Geostatistics. Academic Press, London. 600 pp.

Matheron, G., 1963. Principles of Geostatistics. Economic Geology, Vol. 58, pp 1246-1266.

Olea, R. A., 1984. Systematic Sampling of Spatial Functions. Kansas Geological Survey, Lawrence, Kansas. 57 pp.

Splitstone, D. E., 1987. Sampling Design: Some Very Practical Considerations. (This volume).

Starks, T. H., 1986. Determination of Support in Soil Sampling. Mathematical Geology, Vol. 18, No. 6, pp. 529-537.

Yfantis, E. A., Flatman, G. T., and Behar, J. V., 1987. Efficiency of Kriging Estimation for Square, Triangular, and Hexagonal Grids. Mathematical Geology, Vol. 19, No. 3, pp. 183-205.

Discussion

John Warren
US Environmental Protection Agency

Spatial prediction, as illustrated by these two papers, holds great promise in extracting large amounts of information from situations where extensive data collection is not feasible.

Evan Englund's paper on using kriging in conjunction with the concept of sample support is most interesting and his discussion illustrates clearly the problem of defining "sample" with respect to the physical dimensions of a sample site. The use of "sample support" to avoid confusion with the common statistical understanding of "sample" is to be commended and his definition: "The support on which decisions will be made is rarely the same as that of the samples. The choice of decision support can significantly affect the outcome of an analysis, and should be considered before a sampling program is undertaken." should be included in every Data Quality Objectives (DQO) Program.

The illustration of "sample support" by simulation is particularly useful in demonstrating how the definition of contamination is a function of the particular support chosen. This simulated example should be made part of the Agency's DQO training program as it clearly demonstrates to managers and decision-makers the importance of allocating resources proportional to the ratio between decision-area and sample support area.

Evan's discussion on the choice of most efficient or most practicable sampling design for sampling in a spatial domain is very useful for there does not seem to be a general solution applicable to all sites. Each site must have a unique sampling plan and although the Agency's guidance document, SW-846, endorses simple random sampling, the document does recommend a tailored-to-the-site approach. The revised edition of SW-846 (4th edition with an expected publication date of mid-1990) will provide guidance on most sampling schemes likely to be encountered in spatial statistics, and also gives guidance on the construction of optimal grid sizes for the effective use of kriging.

Noel Cressie's paper is a nice summarization of spatial statistics and introduces some of the scientific notations essential for discussion of kriging. His comparison of trend surface model with the random field model is very useful and clearly illustrates the differences between the two models.

The strength of the paper lies in the excellent discussion of spatial design of networks, Section 3. The problem of adding a monitoring site to an existing network cannot be solved easily and only by making key assumptions on the intent and use of the monitoring data can any solution be attempted. The concept of "average" prediction error (equation 3.4) has much merit as many of the Agency's monitoring networks are intended to measure mean increases/decreases in pollution levels. It seems clear that this notion of optimal network design can be applied to Agency networks where non-statistical considerations in determining actual sites are relatively unimportant.

There is much that needs to be done in order to effectively bring spatial statistics to the practical level at EPA, and these two papers are an important start. There is definitely an interest in spatial statistics as evidenced by the number of people present at the spatial statistics sessions at the annual EPA Conference on Statistics; these two papers should further stimulate this interest.

SAMPLING AND MODELING POLLUTANT PLUMES: METHODS COMBINING DIRECT MEASUREMENTS AND REMOTE SENSING DATA

Ralph E. Folsom
Research Triangle Institute, P.O. Box 12194, Research Triangle Park, NC 27709

1.0 Introduction

The goal of this paper is to describe response or trend surface methods for characterizing groundwater contamination plumes. The work focuses on statistical problems associated with mapping pollutant plumes in the vicinity of hazardous waste (superfund) sites. The importance of conducting low cost screening surveys to obtain indirect measurements of contamination is emphasized. Two such measures of subsurface electrical conductivity are described. To illustrate the utility of such low cost screening measurements, electromagnetics or EM is employed as a covariate in spatial regression models fit to Chloride and Tetrahydrofuran concentrations recorded at a Canadian landfill.

We also explore use of indirect measurement surveys for improving the efficiency of well-site sampling plans. A stochastic regression model provides a flexible theoretical framework for deriving such results. In order to minimize the impact of prediction biases that result from model misspecification, probability sampling and estimation methods are advocated.

In the concluding section, an extended stochastic coefficient model that provides for spatial correlation among the well-site specific intercepts, β_{0i}, is considered. This adaptation establishes a class of spatial regression models encompassing the various linear prediction methods that are collectively referred to as Kriging.

2.0 Response Surface Methods

To set the stage for describing trend surface fitting procedures, a data collection design for a typical hazardous waste site will be formulated. First consider a two dimensional target region, R, symbolizing a specified surface area around a hazardous waste site. Within this region R we let Ω denote a finite set of potential well sites for sampling ground water. These potential well sites can be envisioned as a dense grid of N points or coordinate pairs $w_k = (x_k, y_k)$ blanketing the target region R. By restricting the set of potential well sites to a finite subset of the points in R, we can draw upon the extensively developed methodology of probability sampling. This methodology provides a wide array of randomized site selection plans that can be employed to specify a sample, s, of $n<N$ well locations drawn from Ω.

We will focus attention on sampling plans and associated response surface prediction methods that employ a low cost screening measurement c_k that provides an indirect assessment of ground water contamination at each potential well site in Ω. Application of these covariate based methods assumes that an initial screening survey has been performed to obtain one or more indirect measurements of contamination at a dense set of points Ω in R. Therefore, the extent of the screening survey delimits the set of potential well sites, Ω, for which the necessary covariate data is available. In a subsequent section, two indirect measurements of ground water conductivity are described. One of these methods, namely, Electromagnetics or EM, employs various hand held or vehicle mounted devices that record a continuous set of conductivity related measurements while the operator traverses a transect line across the target region R. For a screening survey employing an EM device, we envision the finite set of potential well sites as points positioned at regular intervals along a set of closely spaced transect lines that fully span the target region.

At each of the n sample well sites $s_i = (x_i, y_i)$, wells are drilled and water samples are collected at various depths. These water samples are transported to a laboratory where the concentrations of various chemical contaminants are measured. In the subsequent formulation of response surface models, the potentially observable concentration of a specific chemical constituent at a particular time point (t_o) and a particular constant depth (d_o) below a potential well site w_k will be denoted by z_k. Given the two-phase data collection scheme outlined above, the goal of our statistical methods will be to predict z_k for all points w_k belonging to Ω. The associated response surface is therefore characterized by a set of N spikes of predicted height $\hat{z}_k$ standing on the dense grid of points Ω that blanket the target region R. In the following section, a notation for the proposed response surface model is presented.

2.1 Response Surface Models

We choose to refer to the class of spatial regression models treated here as 'response surface models' to emphasize their connection to the well developed statistical methodology proposed initially be G.E.P. Box for predicting the output of chemical and industrial processes. These models concentrate on predicting the mean function of the pollution surface. If we denote the potentially observable pollution concentration at the site w_k by z_k, then the stochastic process comprised of the vector of N pollution values z_k is assumed to have a reasonably smooth mean function that depends only on the spatial coordinates (x_k, y_k) and the associated covariate c_k. Formally, this assumption can be stated as follows:

$$E(z_k) = F(x_k, y_k, c_k),$$

for all potential well sites w_k in Ω with the $E(\ )$ operator denoting statistical expectation.

Borrowing from response surface methodology, it is assumed that the function F can be satisfactorily approximated by low order polynomial functions of x, y, and c. It is instructive to note that this is a much more tenable assumption with the covariate c in the polynomial equation than it would be with x and y alone. If one lets X_k denote a (p+1) element row vector containing a leading 1 and p requisite polynomial functions of x, y, and c, then we assume that the expected value of an observable concentration measurement z_k is

$$E(z_k) = X_k\beta \qquad (1)$$

where β is the associated column vector of (p+1) unknown polynomial regression coefficients. The feature that will distinguish our response surface models from the spatial regression models employed in Kriging is the lack of spatial correlation between neighboring z_k values.

Since spatial correlation is induced by microstructure in the concentration surface that is not adequately accounted for by the polynomial mean function in equation (1), having a good indirect measurement c_k to include in X_k will substantially reduce the spatial correlation between residuals $u_k = (z_k - X_k\beta)$ and u_l. The results developed subsequently in this paper are derived from models where the covariance is zero between residuals u_k and u_l that are from different well sites. In the final section extensions of our models are proposed that allow for spatial correlation. These models lead to stochastic coefficient generalizations of 'Universal Kriging' methods. Due to the importance of having a good covariate if simple response surface methods are to be successful, we will devote space in the following section to elaborate on the properties of two such indirect measurements, Electromagnetics and Resistivity.

2.2 Electromagnetics and Resistivity

Electromagnetics devices measure the electrical conductivity of subsurface soil, rock, and groundwater. The specific conductance of the ground water and contaminating pore fluids usually dominates the measurement. The EM measurement responds to significant increases in chloride, sodium, sulfate, potassium, calcium, and magnesium. Variation in these species reflect the contaminant variation at parts per million (ppm) and parts per billion (ppb) levels. EM measurements are made by induction with no ground contact required. The depth of the conic shaped subsurface volume that is assessed by a single EM reading depends on the diameter of the coil that induces the electrical field. The average depth of the conic volumes scanned by EM devices are roughly 1.5, 6, 15, 30 and 60 meters. The two deepest reading devices are vehicle mounted whereas the three shallow scanning instruments are hand held. These EM devices are ideally suited to generating conductivity profiles along lateral transects targeted at one of the six available depths.

A second widely used screening device measures subsurface electrical resistivity or the inverse of conductivity. The Resistivity measuring instrument injects an electrical current into the ground at a pair of electrodes. Like EM, Resistivity measures an average response for a subsurface volume ranging from the surface to the effective depth. The nominal response depth for a particular Resistivity sounding is approximately the electrode spacing. Since the pair of electrodes can in principle be separated by any predesignated distance, the range of possible reading depths is unlimited. Noting that Resistivity measurements require that a pair of electrodes be place in contact with the ground, it is clear that such measurements are slower to obtain than EM readings. For this reason, applications of Resistivity are generally restricted to depth sounding at a limited number of sites. Due to its speed of application for lateral coverage, the EM device is more suitable for the blanket coverage required of our low cost screening survey. In the following section, a data base providing linked EM survey measurements and lab recorded contaminant concentrations is explored. This data base was developed to monitor the magnitude and extent of groundwater contamination around the Gloucester, Ontario landfill.

2.3 The Gloucester Landfill Data

The Gloucester landfill is located six miles south of Ottawa, Ontario. From 1956 to 1973, the landfill was the site of domestic waste disposal. From 1969 to 1980, a special site within the landfill was used to dispose of hazardous university, hospital, and government wastes. By 1985, approximately 60 multiple level and 200 standpipe monitoring wells had been installed. Extensive EM surveys were conducted obtaining indirect conductivity measurements at the six and fifty meter depths. To demonstrate the utility of the EM measurement, homogeneous polynomial models were fitted to 1983 data from 44 wells and to 1985 data from 21 wells.

A degree-p homogeneous polynomials in the x and y spatial coordinates includes all regressors of the form

$$X_{lt} = x_l^r y_l^c$$

such that the polynomial exponents r and c take values zero through p and $r+c \leq p$. Restricting attention to such polynomials is advisable since the associated regression equations produce the same predicted values independent of an arbitrary rigid rotation of the x and y coordinate axis. Since the positioning of x and y coordinate axes over the target region of a hazardous waste site is generally arbitrary, models whose predictions

$$\hat{z}_k = X_k\beta$$

change with the positioning of the coordinate axis should be avoided.

When one is forced by a sparsity of observations at a uniform depth to simultaneously fit concentrations from varying depths, then polynomials that involve the recorded depth d_l are required. This has been the case with the Gloucester data. Since the position of the depth axis is nonarbitrary, models need not be fully homogeneous or invariant under rigid rotations in three dimensions. The covariate measurements c_k are also linked to the potential well sites w_k in an unambiguous fashion that is not altered by repositioning the x and y coordinates.

Table 1 displays a fully homogeneous degree-2 polynomial fit to the natural log of 1983 chloride concentrations recorded at 185 well site by depth combinations. The average log concentration for the 185 measurements is 2.63 which yields 13.87 mg/L on the untransformed scale. With four dimensions among the predictors, the homogeneous degree-2 polynomial has an intercept and fourteen regressors. The multiple R^2 for this model was $R^2 = 0.63$.

The pure error mean square is based on the presence of replicated water samples and associated readings for selected well by depth combinations. The test for lack-of-fit is highly significant. While it is possible that a higher order polynomial would have achieved a nonsignificent lack of fit, we did not attempt any higher order models. Our objective in fitting such models was to demonstrate the utility of the EM covariate (COND50) for a number of different pollutants. To expedite this demonstration, we limited our analysis to the same degree-2 polynomial for all of the pollutant outcomes. The 50 foot EM variate is clearly an important predictor of the 1983 Chloride concentrations. Figures 1 through 3 are included to illustrate a common problem with polynomial regression predictors. Figure 1 presents two so-called bubble plots of the Chloride concentrations observed at 28 wells. The square figures represent 600 meter by 600 meter square regions overlaying the hazardous waste site. The variable size dots denote the size category of the maximum Chloride concentration observed at the indicated location and over the designated depth range. Note from figure 1 that there are only two well locations at x-axis values in excess of 3,000 meters. We have also observed that the 50 meter (EM) conductivity values beyond 3,000 meters on the x axis are all relatively small. In spite of these relatively small values of the EM covariate beyond 3,000 meters, figure 2 shows that the polynomial equation from table 1 predicts a substantial Chloride peak near the center of the 3,100 meter boundary. Figure 3 is a three-dimensional spider-web plot of the predicted values for maximum Chloride concentration in the 15 to 25 meter depth range.

These series of figures illustrate the risk of extending one's mapping region beyond the area where water samples have been analyzed for contaminant concentrations. Another way of addressing this problem is to insure that wells have been located on or near the map's boundaries and at sufficient density within the region to guarantee a reasonable approximation to the surfaces major features. In the following sections, we develop some alternative well location strategies based on different stochastic models for the pollution process.

Before turning to the well site sampling problem, it is instructive to demonstrate that in addition to correlating well with highly ionic contaminants like Chloride, the EM electrical conductance measure can also predict the dispersion of organic contaminants. Table 2 summarizes the fit of our degree-2 model to a set of 41 log concentrations of 1985 Tetrahydrofuran (THF). These concentration measurements were taken at 19 wells and were associated with depths ranging from four to six meters. Even within this narrow depth range, it is clear that the associated variable D is an important predictor of THF concentrations. The six meter EM (COND6) variable is also a strong predictor of THF. Note that in this case the test for lack of fit is nonsignificant. With only nine degrees of freedom for estimating the within site sampling and measurement error variance (PURE ERROR), this nonsignificant result may well be a reflection of poor test power.

3.0 Model Based Purposive Sampling

To define statistical optimality criteria for alternative well samples, we need to specify the variance and covariance function for our stochastic concentration measurements z_k. Recall that the mean function has the general linear form

$$E(z_k) = x_k\beta$$

where x_k denotes a (p+1) element row vector containing a leading 1 and p polynomial functions of the spatial coordinates (x_k,y_k) and the covariate c_k. In our original formulation of this problem, the subscripts-k were linked to potential well sites w_k corresponding to points (x_k,y_k) blanketing the region R that was to be mapped. Recall that the dense grid of N potential well sites where covariate measurements c_k had been obtained was denoted by Ω.

Implicit in this design formulation is the notion that all wells will be constructed to sample at the same set of depths prespecified on the basis of geophysical and hydrogeological considerations. To account for the possibility of having multiple depth targeted covariates like our six meter and fifty meter EM measurements, we can allow x_k to include the full vector of covariates, say $\underset{\sim}{c}_k$, so that the mean function has the same linear form for each depth level-d; that is,

$$E(z_{kd}) = x_k\beta_d$$

for each of the prespecified depths-d.

Given a common linear form for the mean function at each depth-d and the further assumption that the z_{kd} are independent with common variance σ_d^2, one observes that there is a single sample-s of n potential well sites chosen from Ω that is simultaneously optimal for mapping pollutant concentrations z at all depths-d. To specify an optimum sample-s, we require an optimality criterion. Under our response surface model, the best linear unbiased estimator for a future value of z_{kd} is

$$\hat{z}_{skd} \equiv x_k b_{sd}$$

where b_{sd} is the ordinary least squares (OLS) estimator for the vector of depth specific polynomial regression coefficients β_d. The variance of this best linear predictor is

$$\mathrm{Var}(\hat{z}_{skd}) = (x_k W_s^{-1} x_k^T)\sigma_d^2 \equiv \nu_{sk}\sigma_d^2 \qquad (2)$$

where W_s^{-1} is the inverse of the left-hand-sides matrix in the normal equations for estimating b_s, and the superscript T denotes matrix transposition. In response surface parlance, W_s is the incidence matrix of the design-s. If X_s denotes the n by (p+1)matrix whose rows correspond to the n vectors x_k associated with the sample well sites-s, then the sample incidence matrix is

$$W_s \equiv (X_s^T X_s).$$

If we further let the measurements $\tilde{z}_{kd}$ denote a

set of future values for the pollutant concentrations at depth-d, then the squared euclidean distance between the vector $\tilde{z}_\Omega(d)$ of N future values and our vector $\hat{z}_\Omega(sd)$ of best linear predictors can be depicted as follows:

$$L_\Omega(sd) \equiv \sum_{k=1}^{N} (\tilde{z}_{kd} - \hat{z}_{skd})^2. \quad (3)$$

We will define the optimum sample-s as that particular sample of size n that minimizes the expected squared distance between the predicted surface $\hat{z}_\Omega(sd)$ and a future independent realization of the contamination process $\tilde{z}_\Omega(d)$.

With $L_\Omega(sd)$ in equation (3) representing our loss function, the associated expected loss or risk function is

$$R_\Omega(sd) \equiv \sum_{k=1}^{N} E(\tilde{z}_{jd} - \hat{z}_{skd})^2. \quad (4)$$

This risk function is proportional to the average mean-squared prediction error where the averaging extends over all potential well sites w_k belonging to Ω; that is, over all N points where the z_d surface is to be mapped. The independent residuals assumption of our Response Surface model leads to the risk function

$$R_\Omega(sd) = \sum_{k=1}^{N} (\sigma_d^2 + \nu_{sk}\sigma_d^2) = N\sigma_d^2 + \sigma_d^2 \left[\sum_{k=1}^{N} \nu_{sk} \right]. \quad (5)$$

Now, one observes that the sample-s that minimizes (5) must also minimize

$$\sum_{k=1}^{N} \nu_{sk} = tr(W_\Omega W_s^{-1}) \quad (6)$$

where W_Ω is the universe level incidence matrix based on all N potential well sites and tr( ) is the trace operator. An equivalent criterion that is proportional to (6) replaces W_Ω and W_s by the associated mean square and cross product matrices $\bar{W}_\Omega \equiv (W_\Omega + N)$ and $\bar{W}_s \equiv (W_s + n)$.

In terms of these matrices, the criterion in equation (6) reduces to

$$\sum_{k=1}^{N} \nu_{sk} = (N + n)tr(\bar{W}_\Omega \bar{W}_s^{-1}) . \quad (7)$$

To simplify this criterion further, one can perform an orthonormal transformation on the X_k vectors so that at the universe level the transformed mean incidence matrix is the identity matrix. Subsequent to this transformation, the optimality criterion simplifies to

$$\sum_{k=1}^{N} \nu_{sk} = (N + n)tr(\bar{W}_s^{-1}) . \quad (8)$$

If we assume further that the potential well sites w_k in Ω were positioned at the nodes of a regular square grid, then the original x_k and y_k well coordinates are orthogonal to each other. In this case the transformed coordinates are simply the standardized versions, say x_k and y_k, formed by subtracting their respective universe-Ω level means and dividing the resulting deviations by the associated root mean squared deviations.

To illustrate the nature of samples that minimize our risk function, consider a simple degree-1 polynomial function. A response surface model of degree-1 and a mapping universe Ω of points (x_k,y_k) positioned on a regular square grid yields the following orthonormalized vector of linear predictors

$$\chi_k \equiv (1, x_k, y_k, \ell_k)$$

where the transformed x_k and y_k coordinates are the standardized versions described above and ℓ_k denotes the orthonormal version of the covariate c_k. The transformed covariate has the form

$$\ell_k \equiv (c_k - \hat{c}_k) + \sigma_{\ell\Omega} \quad (9)$$

where $\hat{c}_k$ is the Ω level ordinary least squares predictor for c_k based on the first three elements of χ_k. The divisor $\sigma_{\ell\Omega}$ is the square root of the average squared residual $(c_k - \hat{c}_k)^2$.

Returning to the optimality criterion in equation (8), one observes that for samples-s that are orthogonal in the sense that $\bar{W}_s$ is diagonal, the risk function is proportional to

$$tr(\bar{W}_s^{-1}) = 1 + \sigma_{xs}^{-2} + \sigma_{ys}^{-2} + \sigma_{\ell s}^{-2} \quad (10)$$

where the σ_s^2 quantities in equation (10) are the sample averages of the squared variates x_k^2, y_k^2, and ℓ_k^2. For such orthogonal samples, it is clear that our risk function is minimized by choosing points w_k with the largest values of x_k^2, y_k^2, and ℓ_k^2. Such points lie on or near the spatial boundary of Ω where large positive or negative residuals ℓ_k are observed. Note that these are points where the degree-1 spatial polynomial predicts the covariate poorly.

Of course, sample orthogonality is not a prerequisite for minimizing tr $(\bar{W}_s^{-1})$. Focusing on the subset of orthogonal samples was merely an expedient used to obtain a simple characterization of designs that minimize $tr(\bar{W}_s^{-1})$. Various iterative computer algorithms have been devised to find near optimal samples based on minimizing criteria like equations (7) and (8). Mitchell (1974) presents one such algorithm.

It is instructive to note that minimizing equation (6) amounts to minimizing the average variance of the predicted values $\hat{z}_{skd}$ over Ω assuming that the form of the mean function is known. Robust response surface methods have also been developed that seek to minimize the average mean squared prediction error of equation (4) assuming that the true mean function for the process is a higher degree

polynomial than that being used to produce the $\hat{z}$ predicted values. In this case, the dominant term in the risk function of equation (4) is the sum over Ω of the squared prediction biases

$$B_{skd} = E(\tilde{z}_{kd} - \hat{z}_{skd}).$$

The robust designs minimize the average squared prediction bias and therefore the average mean squared prediction error by choosing samples where the associated incidence matrix Ω_s equals or closely approximates the population incidence matrix Ω_Ω. In this context, it has been pointed out that probability samples yield incidence matrices Ω_s that equal Ω_Ω on average over repeated samples and that come to closely approximate Ω_Ω as the sample size becomes sufficiently large.

In the following section, we develop probability sample designs that attempt to minimize variance by over representing boundary points where large ℓ_k residuals occur. To minimize squared prediction bias, we advocate inverse sample inclusion probability weighted estimates of the regression coefficients.

4.0 Probability Sampling for Response Surfaces

To develop a satisfactory probability sampling strategy for response surfaces, one must satisfy several requirements. These requirements include:

1.) An explicitly defined finite population of potential well sites (Ω).
2.) Sensible finite population parameter definitions. For example, define the vector of regression coefficients, say β_Ω, that would result from selecting and analyzing the full population of N potential well sites.
3.) Efficient probability sample design. This requirement calls for a specification of well site selection probabilities
$$\pi_{sk} = \text{Prob}(w_k \epsilon s)$$
such that the variance of our predicted values will be minimized.
4.) An asymptotically design unbiased (ADU) estimator, say β_s, for the finite population parameter β_Ω and an ADU estimator for the sampling covariance matrix of β_s. The ADU property implies that under suitable regularity conditions the bias of a statistic over repeated probability samples tends to zero as the sample size n increases.

Our previous discussion has anticipated the need for an explicitly defined finite population. In the subsequent development, a stochastic coefficient model is employed to address requirements 2.) through 4.) above.

4.1 The Stochastic Coefficient Model

In our application of stochastic coefficient models, each of the n sample well sites w_i is allowed to have its own vector of regression parameters β_i. Also providing for multiple water samples and associated lab determinations at depth-d_0, the model becomes

$$z_{ij} = X_i\beta_i + e_{ij}. \qquad (11)$$

In equation (11), the errors e_{ij} are assumed to be independent and identically distributed random variables with common mean zero and variance σ^2. These errors are linked to within well variation between the water samples-j and measurement variation in the associated m_i lab determinations. As before, X_i is the vector of (p+1) polynomial regressors based on x_i, y_i, and ζ_i. The stochastic coefficient vectors β_i are similarly assumed to be independent and identically distributed with common mean vector γ and covariance matrix Δ.

Alternatively, one can recast equation (11) in terms of a so called mixed model where the (p+1) vector of expected regression coefficients γ represent the fixed effects and the vectors $\eta_i \equiv (\beta_i - \gamma)$ of deviations represent the random effects. This mixed model form is

$$z_{ij} = X_i\gamma + X_i\eta_i + e_{ij} \qquad (12)$$

The second term in equation (12) represents the lack-of-fit contribution to the residual. Now, if we let $\bar{z}_i$ denote the mean of our m_i measurements at well-i and assume further that the η_i and e_{ij} errors are independent, then

$$\text{Var}(\bar{z}_i) = [X_i\Delta X_i^T + \sigma^2/m_i] \equiv V_i^2. \qquad (13)$$

To specify the model based estimator for γ, the fixed regression coefficients, we let $\bar{z}_s$ denote the (nx1) column vector of mean concentrations and define V_s as the (nxn) diagonal covariance matrix with diagonal elements V_i^2.

In terms of our mixed model notation, the best linear unbiased (BLU) estimator for the fixed coefficient vector γ given that Δ and σ^2 were known is

$$\gamma_s = (X_s V_s^{-1} X_s^T)^{-1} X_s^T V_s^{-1} \bar{z}_s . \qquad (14)$$

Recall that X_s is the n by (p+1) matrix comprised of the row vectors X_i. The individual well site level coefficient vectors have the following BLU estimator when Δ and σ^2 are assumed known:

$$\beta_{sk} = \begin{cases} \gamma_s + \Delta X_k^T (\bar{z}_k - X_k\gamma_s) + V_k^2 & \text{for all } k \epsilon s \\ \gamma_s & \text{for all } k \notin s \end{cases} \qquad (15)$$

Note that for the sample well sites ($k\epsilon s$) the β_{sk} estimator has a second term, say η_{sk}, that estimates the vector of random effects η_k.

The associated set of BLU predictors for the z surface have the form

$$\hat{z}_{sk} = X_k\beta_{sk} = \begin{cases} X_k\gamma_s + X_k\eta_{sk} & \text{for all } k \epsilon s \\ X_k\gamma_s & \text{for all } k \notin s \end{cases} \qquad (16)$$

The stochastic coefficient model results in a lack-of-fit adjustment to the $\hat{z}$ predictors for wells w_k that belong to the sample-s. When the measurement variance contribution in V_k^2, namely $\sigma^2 + m_k$, is negligible relative to the lack-of-fit contribution ($X_k\Delta X_k^T$) then $\hat{z}_{sk} = \bar{z}_k$ for the sample well sites. When the measurement variance is not negligible, then $\hat{z}_{sk}$ for $k\epsilon s$ has the form of a James-Stein shrinkage estimator; that is, for $k\epsilon s$

$$\hat{z}_{sk} = a_k \bar{z}_k + (1-a_k)(X_k \gamma_s)$$

where the weight a_k on $\bar{z}_k$ is the fraction of the variance V_k^2 that is accounted for by the lack-of-fit contribution $(X_k \Delta X_k^T)$.

While the set of predictors in equation (16) are best in the sense of minimum mean squared prediction error over Ω, the associated surface is discontinuous at the sample well sites. To produce an appealing graphic representation of the surface, a smoother prediction function may be desirable. For this purpose one can derive a common vector of regression coefficients to substitute for the site specific β_i vectors, say β_{s*}, such that the set of smooth linear predictors

$$z^*_{sk} \equiv X_k \beta_{s*} \qquad (17)$$

minimize the risk function in equation (4) with the z^*_{sk} predictors replacing z. In this instance, one minimizes the risk function $R_\Omega(\beta_{s*})$ with respect to the elements of β_{s*} given the sample-s. Note also that the expectation defining $R_\Omega(\beta_{s*})$ is in terms of our stochastic coefficient model.

Pfefferman (1984) shows that with Δ and σ^2 known, the β_{s*} that minimizes the average mean squared prediction error has the form:

$$\beta_{s*} = \left[\sum_{k=1}^{N} X_k^T X_k\right]^{-1} \left[\sum_{k=1}^{N} (X_k^T X_k)\ \beta_{sk}\right] \qquad (18)$$

where the β_{sk} are as defined in equation (15). Folsom (1985) shows that when the number of replicate water samples m_i per site is a constant $\bar{m}$, β_{s*} can be recast in terms of the generalized least squares estimator γ_s of equation (14) and the ordinary least squares (OLS) estimator b_s. Letting $\bar{R}_s = (X_s^T \bar{z}_s) \div n$ denote the vector of mean cross-products in the right-hand sides of the OLS estimation equations, we have

$$\beta_{s*} = \bar{W}_\Omega^{-1}[\bar{R}_s - (\bar{W}_s - \bar{W}_\Omega)\gamma_s - (1-f)\bar{W}_s(b_s - \gamma_s)] \qquad (19)$$

where $f = n \div N$ is the sampling fraction.

Recall that $\bar{W}_s$ and $\bar{W}_\Omega$ denote the mean incidence matrices for the sample-s and the universe-Ω respectively.

Equation (19) shows that when one has wells at all potential sites and $s = \Omega$ then the minimum risk common β is the universe level OLS estimator

$$\beta_\Omega = \bar{W}_\Omega^{-1} \bar{R}_\Omega \qquad (20)$$

with $\bar{R}_\Omega \equiv (X^T \bar{z}_\Omega) \div N$ denoting the population analog of $\bar{R}_s$. Alternatively, as the sampling fraction $f = (n \div N)$ approaches zero the minimum risk common β tends to the generalized least squares (GLS) estimator γ_s. It is also interesting to note that when the sample incidence matrix $\bar{W}_s$ matches the universe level matrix $\bar{W}_\Omega$ then the best common β simplifies to $\beta_{s*} = \gamma_s + f(b_s - \gamma_s)$.

The fact that our best common coefficient vector β_{s*} goes to β_Ω as the sampling rate f goes to 1 suggests that an appropriate definition for the finite population β parameter is β_Ω. In the following section, we propose an inverse selection probability weighted analog of the OLS estimator b_s, namely b_π, that is asymptotically design unbiased for β_Ω. In addition to this desirable large sample property, we propose a sample design with inclusion probabilities π_i proportional to the residual variances V_i^2 dictated by a simplified form of our stochastic coefficient model. When this simplified model holds, our inverse π weighted estimator b_π is equivalent to the best small sample estimator γ_s.

4.2 Efficient Probability Sampling

Having settled on β_Ω as our finite population parameter, an ADU estimator for β_Ω is

$$b_\pi \equiv (X_s^T \pi_s^{-1} X_s)^{-1}(X_s^T \pi_s^{-1} \bar{z}_s) \qquad (21)$$

where π_s is the (nxn)) diagonal matrix with diagonal elements π_{si} corresponding to the inclusion probabilities for the well sites w_i that belong to the sample-s. In order for b_π to be close to the optimum estimator β_{s*} which approaches γ_s when the sampling rate f is small, we require a sample design with the well site selection probabilities π_{si} proportional to our model residual variance; that is,

$$\pi_{si} = \theta V_i^2 = \theta[(X_i \Delta X_i^T) + (\sigma^2 + \bar{m})] \qquad (22)$$

where θ is a proportionality constant and $\bar{m}$ is the common number of water samples proposed for each sample well site.

The set of selection probabilities specified in equation (22) are clearly functions of the unknown model parameters Δ and σ^2. For equation (22) to be a useful design prescription, we must either have some reasonably good prior estimates, say Δ_0 and σ_0^2, of the model parameters or we must simplify the model sufficiently so that V_i^2 is proportional to a known quantity. To accomplish the latter simplification, we consider the special case where the elements of the stochastic coefficient vector β_i are assumed to be uncorrelated and to have common variance δ^2. In this case the stochastic coefficient covariance matrix Δ reduces to the scalar δ^2 multiplied by the identity matrix $I_{(p+1)}$.

If we also assume that the measurement variance contribution $(\sigma^2 + \bar{m})$ to V_i^2 is negligible relative to the stochastic coefficient contribution, then equation (22) reduces to

$$\tilde{\pi}_{si} = \theta \tilde{V}_i^2 = \theta\delta^2(X_i X_i^T) \ . \qquad (23)$$

Note that in equation (23), the selection probabilities are proportional to the known squared lengths $L_i^2 = (X_i X_i^T)$ of the predictor vectors X_i.

In order to make the simplified variance model in equation (23) more plausible, we recommend the orthonormal transformation of section 3.0. Recall that this transformation

yields new predictor vectors χ_i with the property that the associated universe level mean incidence matrix $\bar{W}$ reduces to $I_{(p+1)}$. In addition to its potential for rendering the β_{it} stochastic coefficients independent and equivariant, this orthonormal transformation leads to an interesting formal justification for the inclusion probabilities in equation (23). For our simplified model, we can show that equation (23) provides the vector of selection probabilities $\pi_\Omega \equiv (\pi_1, \ldots, \pi_\Omega)$ that minimize the risk function

$$R(\underline{\pi})_\Omega = E_\xi E_s \left\{ \sum_{k=1}^{N} (\tilde{z}_k - \chi_k b_\pi)^2 + N \right\} \qquad (24)$$

where E_s denotes expectation over repeated probability samples-s and E_ξ expectation over the stochastic coefficient model. The values of π_i that minimize (24) are

$$\pi_i = n\left\{ L_i V_i + \left[\sum_{k=1}^{N} L_k V_k \right] \right\} . \qquad (25)$$

Recall, that $L_i^2 = (\chi_i \chi_i^T)$ and V_i^2 is the model variance of $\bar{z}$. In the simplified model where V_i^2 is proportional to L_i^2, we note that the optimum π_i are also proportional to L_i^2. Therefore, under the simplified model, the well site selection probabilities that minimize the risk function of equation (24) are also the probabilities that render our ADU estimator b_π equivalent to the model-best small sample estimator γ_s.

The probability sampling scheme outlined above had intuitive appeal in light of the model based purposive sampling results of section 3.0. Given that our well site selection probabilities are proportional to the squared length of the associated vectors χ_i of orthonormalized predictors, we observe that the longest χ_i vectors are associated with sites on or near the boundary of the sampling region where atypically large values of the transformed covariate are observed. Recall that the transformed covariate for site w_k is proportional to $(c_k - \hat{c}_k)$, the difference between the observed covariate and its (x,y) coordinate polynomial predictor. It is clear that large $(c_k - \hat{c}_k)$ residuals will occur where abrupt spikes or holes are observed in the covariate surface. Near the sampling region boundary it will be the spikes that generate large residuals and correspondingly large site selection probabilities.

The model based purposive sample designs of section 3.0 focused exclusively on these boundary region sites where atypically large covariate values were observed. Our randomized sampling design assigns the highest selection probabilities to those points while insuring that all potential sites have a nonzero chance of selection. As indicted in section 3.0, this defining property of probability samples ensures that the inverse π weighted sample incidence matrix $\bar{W}_\pi$ equals $\bar{W}_\Omega$ on average over repeated samples and comes to closely approximate $\bar{W}_\Omega$ as the sample size becomes sufficiently large. Therefore, in the context of response surface prediction, we see that probability samples can be both <u>efficient</u> from the standpoint of controlling the variance of prediction errors and <u>robust</u> in terms of controlling the average squared prediction bias.

4.3 <u>Design Consistent Estimation</u>

In this section, we consider a design consistent adaptation of the best common β as defined by equation (19). Recall that our orthogonal transformation of the predictor vectors reduces the universe level mean incidence matrix to the identity matrix. With this transformation, the best common β reduces to

$$\beta_{s*} = [\bar{R}_s - (\bar{W}_s - I_{p+1})\gamma_s - (1-f)\bar{W}_s(b_s - \gamma_s)]. \qquad (25)$$

Examining equation (25), we observed that the two leading terms would yield an ADU estimator for $\beta_\Omega = \bar{R}_\Omega$ if the inverse π-weighted estimators $\bar{R}_\pi$ and $\bar{W}_\pi$ were substituted for $\bar{R}_s$ and $\bar{W}_s$.

Specifically, we define the design unbiased estimators

$$\bar{R}_\pi \equiv \sum_{i \in s} \chi_i^T \bar{z}_i + N\pi_i$$

and

$$\bar{W}_\pi \equiv \sum_{i \in s} \chi_i^T \chi_i + N\pi_i .$$

Making the indicated substitutions, we obtain the following ADU estimator for β_Ω:

$$\beta_s \equiv [\bar{R}_\pi - (\bar{W}_\pi - I_{p+1})\gamma_s] . \qquad (26)$$

The estimator in (26) is clearly design unbiased in large samples since $\bar{W}_\pi \to \bar{W}_\Omega = I_{p+1}$ and $\bar{R}_\pi \to \bar{R}_\Omega = \beta_\Omega$. The form of β_s in equation (26) reveals how knowledge of the population level mean incidence matrix $\bar{W}_\Omega = I_{p+1}$ is employed to adjust the unbiased estimator $\bar{R}_\pi$ for the observed sampling deviation $(\bar{W}_\pi - \bar{W}_\Omega)$. In this form, β_s has the appearance of a vector valued regression estimator with coefficients drawn from the generalized least squares (GLS) estimator γ_s.

To form the GLS estimator (γ_s) employed in (26), we require a sample estimator for the stochastic coefficient covariance matrix Δ and the measurement variance component σ^2. We are inclined to reduce Δ to a diagonal matrix and apply restricted maximum likelihood (REML) methods to obtain the sample estimators Δ_s and σ_s^2. In terms of these statistics, we estimate the variance of the site-i mean $\bar{z}_i$ as

$$V_{si}^2 \equiv [\chi_i \Delta_s \chi_i^T + (\sigma_s^2 + m_i)].$$

Our GLS estimator γ_s is obtained from the associated diagonal covariance matrix. An alternative form for β_s emphasizes this GLS component. Note that for a general set of untransformed predictors, the two leading terms of equation (19) lead to the ADU adaptation

$$\beta_s \equiv \gamma_s + \bar{W}_\Omega^{-1} \bar{W}_\pi (b_\pi - \gamma_s) . \qquad (27)$$

In this form, the second term performs a bias correction when the mean function is not

adequately approximated by the specified (p+1) variable linear predictor. In this case, the correction renders β_S asymptotically design unbiased for β_Ω, the universe level OLS coefficients.

Recall that the original best common β estimator β_{S*} of equation (19) also converged to β_Ω as the sample-s grew to encompass the entire population of potential sites. When our simplified variance model holds approximately with $\Delta = \delta^2 I_{(p+1)}$ and $(\sigma_s^2 + m_i)$ negligible, then the inverse p-weighted estimator $b_\pi \equiv \bar{W}_\pi^{-1}\bar{R}_\pi$ will closely approximate the model based GLS estimator γ_S. In this situation, the degree of correction will be slight and β_S will closely approximate γ_S.

In addition to combining the efficiency of the GLS coefficient vector γ_S and the robustness of the π-weighted analog b_π, the regression estimator β_S makes explicit use of our knowledge of the population matrix $\bar{W}_\Omega$. Effective use of this added information should improve response surface prediction. The following section relates an interesting connection between our simplified stochastic coefficient model and a robust covariance matrix estimator for b_π.

4.4 Robust Covariance Matrix Estimation

In a recent paper, C.F.J. Wu (1986) recommends a weighted jackknife covariance matrix estimator that is strictly unbiased when the model assumptions hold. Wu's version of the delete-one jackknife has the added robustness property of asymptotic unbiasedness when the variance model does not hold. For this robustness property to hold requires only that the residuals be uncorrelated. Recall that the variance model motivating our π-weighted vector of regression coefficients has

$$V_i^2 = \delta^2(\chi_i\chi_i^T) \equiv \delta^2 L_i^2 .$$

The associated site selection probability is

$$\pi_i = n\, L_i^2 + N(p+1).$$

As applied to our p-weighted coefficient vector b_π, Wu's jackknife covariance matrix estimator $cov_{Wu}(b_\pi)$ has the form

$$[(p+1)^2+n]\, \bar{W}_\pi^{-1}\left\{\sum_{i=1}^{n}\chi_i^T\chi_i r_i^2 + L_i^4[n-(p+1)\xi_i]\right\}\bar{W}_\pi^{-1} \quad (28)$$

where the r_i quantities depict the observed mean residuals

$$r_i \equiv (\bar{z}_i - \chi_i b_\pi).$$

The ξ_i quantity in the divisor of equation (28) is defined as:

$$\xi_i \equiv [(\chi_i\bar{W}_\pi^{-1}\chi_i^T) + (\chi_i\bar{W}_\Omega^{-1}\chi_i^T)]. \quad (29)$$

Recall that with the χ_i orthonormalized, the universe level mean incidence matrix $\bar{W}_\Omega = I_{(p+1)}$.

Wu's results imply that when $V_i^2 = \delta^2 L_i^2$, then

$$nr_i^2 + [n-(p+1)\xi_i] = \hat{V}_i^2 \quad (30)$$

is a model unbiased estimator for V_i^2 and (28) is strictly model unbiased. On the other hand, when $V_i^2 \neq \delta^2 L_i^2$ the estimator in (28) is still asymptotically unbiased as long as the unobserved residuals

$$u_i = (\bar{z}_i - \chi_i\gamma)$$

are uncorrelated. This result follows since $\hat{V}_i^2$ is asymptotically unbiased.

It is interesting to note that when the π-weighted incidence matrix $\bar{W}_\pi$ matches the universe level analog $\bar{W}_\Omega = I_{(p+1)}$, then $\xi_i = 1$ for all sites-i and the bias correction factor applied to the residuals in equation (30) reduces to the familiar degrees-of-freedom adjustment. It is also instructive to note that in the case of pps with replacement sampling, the estimator with $\xi_i = 1$ reduces to Fuller's (1974) version of the Taylor Series or linearization approximation for the probability sampling covariance matrix. The $\xi_i = 1$ version is also analogous to Hinkley's (1977) weighted jackknife.

Folsom's (1974) version of the linearization estimator employed the correction factor $[n+(n-1)]$. Wu also shows that the standard bootstrap covariance estimator for b_π can be approximated by equation (28) with the bias correction factor totally removed. While all of these variations are asymptotically equivalent, Wu's simulation results suggest that the bias correction of equation (28) can be important for small samples.

The connection between these robust covariance matrix estimators and our stochastic coefficient model follows from the definition of the site specific coefficient estimators

$$b_{\pi i} \equiv b_\pi + \Delta_o\chi_i^T(\bar{z}_i - \chi_i b_\pi) + V_{io}^2 \quad (31)$$

where $\Delta_o = \delta^2 I_{(p+1)}$ and $V_{io}^2 = \delta^2 L_i^2$. As defined in equation (31) one observes that the $b_{\pi i}$ coefficients sum to zero across sites. Given this definition of $b_{\pi i}$, we can recast Wu's covariance matrix estimator in terms of the associated sample covariance matrix

$$S_b \equiv \sum_{i=1}^{n}(b_{\pi i} - b_\pi)(b_{\pi i} - b_\pi)^T + [n-(p+1)\xi_i] .$$

Employing the matrix S_b, the robust covariance matrix estimator becomes

$$cov_{Wu}(b_\pi) = (p+1)^2(\bar{W}_\pi^{-1}S_b\bar{W}_\pi^{-1}) + n . \quad (32)$$

Analogous covariance matrix estimators can be specified for our GLS coefficient vector γ_S and the design consistent adaptation $\beta_S \equiv \gamma_S + \bar{W}_\Omega^{-1}\bar{W}_\pi(b_\pi - \gamma_S)$. We believe that these model bias corrected versions of the jackknife/linearization estimators warrant further evaluation.

5.0 Summary Comments

In summary, we have made a case for the utility of remote sensing data, particularly electromagnetics (EM), as a predictor of ground water contamination levels. Stochastic coefficient models were employed to motivate interesting probability sampling alternatives to model based designs. We believe that the good large sample properties of probability sampling methods merit further examination in the moderate sample size arena. In this context, we favor coefficient estimators that are efficient when the model holds and asymptotically design unbiased when the model fails. An interesting jackknife covariance matrix estimator incorporating a model bias correction was also introduced.

In the area of further research, we are currently exploring an extended class of stochastic coefficient models that permit the regression intercepts β_{0i} to exhibit spatial autocorrelation. Restricted maximum likelihood (REML) methods are being employed to estimate the covariance function parameters. Our goal is to develop efficient and robust statistical procedures that combine the best features of 'response surface' methods for mean function prediction and "Kriging" procedures for fitting spatial covariance functions. REML estimation is attractive in this context since it combines model based efficiency in the gausian environment with a known robustness property in the nongausian case. For linear covariance function parameters, REML is equivalent to itterated MINQUE (minimum norm quadratic unbiased estimation).

6.0 References

Folsom, R. E., (1974). National Assessment Approach to Sampling Error Estimation. Sampling Error Monograph. RTI final report prepared for the Education Commission of the States' National Assessment project.

Folsom, R. E., C. A. Clayton, F. J. Potter, and R. M. Lucas, (1985). The Use of Geophysical Techniques in the Statistical Estimation of Magnitude and Extent for Subsurface Contamination at Hazardous Waste Sites. RTI final report prepared for EPA's Statistical Policy Branch.

Fuller, W. A., (1975). Regression Analysis for Sample Surveys. Sankhyā C 37, 117-132.

Hinkley, D. V., (1977). "Jackknifing in Unbalanced Situations." Technometrics, 19, 285-292.

Mitchell, T. J., (1974). "An Algorithm for the Construction of 'D-Optimal' Experimental Designs," Technometrics, 16, 203-210.

Pfeffermann, D., (1984). "On Extensions of the Gauss-Markov Theorem to the Case of Stochastic Regression Coefficients." Journal of the Royal Statistical Society, 46, No. 1, 139-148.

Wu, C. F. J., (1986). "Jackknife, Bootstrap and other Resampling Methods in Regression Analysis." The Annals of Statistics, 14, 1261-1343.

Table 1: Degree-2 Polynomial Fit to 1983 Log-Chloride

Regression	DF	Type I SS	R-Square	F-Ratio	Prob
Linear	4	87.58111894	0.2159	24.79	0.0001
Quadratic	4	90.57004964	0.2232	25.64	0.0001
Crossproduct	6	77.41551896	0.1908	14.61	0.0001
Total Regress	14	255.56669	0.6299	20.67	0.0001
Residual	**DF**	**SS**	**Mean Square**	**F-Ratio**	**Prob**
Lack of Fit	137	141.12041	1.03007597	3.768	0.0001
Pure Error	33	9.02196057	0.27339274		
Total Error	170	150.14237	0.88319040		

Parameter	DF	Estimate	Std Dev	T-Ratio	Prob
Intercept	1	-15.85168215	4.41599408	-3.59	0.0004
X	1	0.02120588	0.009779667	2.18	0.0307
Y	1	-0.14238897	0.01579998	-9.01	0.0001
D	1	0.29552937	0.30227869	0.98	0.3296
Cond50	1	1.34988504	0.39302999	3.43	0.0007
X*X	1	-0.000043146	0.000012153	-3.55	0.0005
Y*X	1	0.000215274	0.000032573	6.61	0.0001
Y*Y	1	-0.000275732	0.000025697	-10.73	0.0001
D*X	1	0.001216132	0.000334634	3.63	0.0004
D*Y	1	0.001437620	0.000537719	2.67	0.0082
D*D	1	-0.009610226	0.007018137	-1.37	0.1727
Cond50*X	1	0.000273421	0.000723282	0.38	0.7059
Cond50*Y	1	0.003379166	0.000789312	4.28	0.0001
Cond50*D	1	0.02522552	0.01375624	1.83	0.0684
Cond50*Cond50	1	-0.06448431	0.01542040	-4.18	0.0001

Table 2. Degree-2 Polynomial Fit to 1985 Log-Tetrahydrofuran

REGRESSION	DF	TYPE I SS	R-SQUARE	F-RATIO	PROB
LINEAR	4	10.13806033	0.3761	9.96	0.0001
QUADRATIC	4	2.27126830	0.0842	2.23	0.0931
CROSSPRODUCT	6	7.93565507	0.2944	5.20	0.0012
TOTAL REGRESS	14	20.34498370	0.7547	5.71	0.0001
RESIDUAL	**DF**	**SS**	**MEAN SQUARE**	**F-RATIO**	**PROB**
LACK OF FIT	17	4.64649597	0.27332329	1.250	0.3781
PURE ERROR	9	1.96751001	0.21861222		
TOTAL ERROR	26	6.61400598	0.25438485		

PARAMETER	DF	ESTIMATE	STD DEV	T-RATIO	PROB
INTERCEPT	1	-58.56598111	27.55493712	-2.13	0.0432
X	1	0.10786243	0.06752790	1.60	0.1223
Y	1	0.21557419	0.15441818	1.40	0.1745
D	1	-8.21549669	2.26361317	-3.63	0.0012
COND6	1	18.69031289	6.05457900	3.09	0.0048
X*X	1	-0.000029699	0.000179072	-0.17	0.8696
Y*X	1	0.001158774	0.001347590	0.86	0.3977
Y*Y	1	-0.000902490	0.001604787	-0.56	0.5787
D*X	1	0.02107810	0.005249011	4.02	0.0005
D*Y	1	-0.006061689	0.01139379	-0.53	0.5992
D*D	1	0.35457561	0.13182192	2.69	0.0123
COND6*X	1	-0.04542798	0.01283282	-3.54	0.0015
COND6*Y	1	-0.03727344	0.01779248	-2.09	0.0461
COND6*D	1	0.23862751	0.13736266	1.74	0.0942
COND6*COND6	1	-0.61796587	0.23241860	-2.66	0.0132

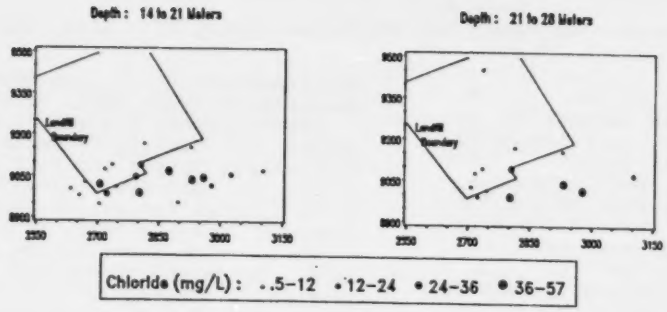


Figure 1. Distribution of 1983 Chloride Concentration by Depth of Well

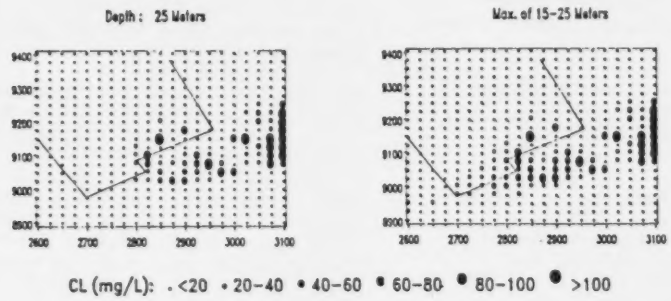


Figure 2. Predicted Values of 1983 Chloride for Deep Depths

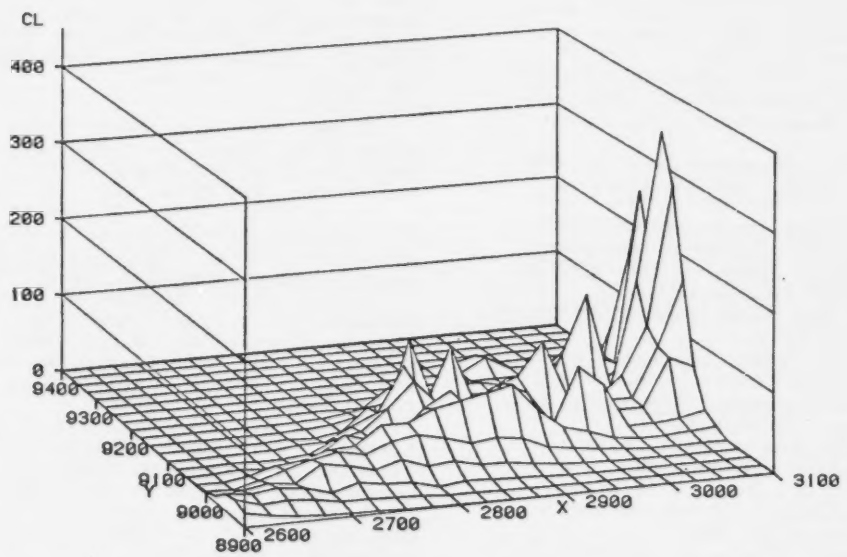


Figure 3. Predicted Values of 1983 Chloride (mg/L)
Maximum Concentration at 15-25 Meters

"VALIDATION" OF AIR POLLUTION DISPERSION MODELS

Anthony D. Thrall
Electric Power Research Institute, 3412 Hillview Avenue, Palo Alto, CA 94303

ABSTRACT

Computer simulations of the dispersion of pollutants through the air are used extensively to guide regulatory decisions concerning air quality. There are a variety of air dispersion models, some of which are classified by the U.S. EPA as "preferred" for certain types of applications, others of which are classified as "alternatives," and the remainder of which must prove themselves "valid" for regulatory application according to a specified evaluation protocol. The criteria, evaluation procedures, and test results are reviewed for the case of Gaussian plume models of sulfur dioxide emissions from isolated point-sources, such as coal-fired power plants located in rural areas.

REFERENCES

American Meteorological Society, 1983. Uncertainty in Air Quality Modelling. Report of a Workshop at Woods Hole, MA, September 1982. AMS, Boston, MA.

Bowne, N.E., Londergan, R.J., Murray, D.R. and Borenstein, H.S., 1983. Overview, Results, and Conclusions for the EPRI Plume Model Validation and Development Project: Plains Site. EA-3074, the Electric Power Research Institute, Palo Alto, CA.

Bowne, N.E., Londergan, R.J. and Murray, D.R., 1985. Summary of Results and Conclusions for the EPRI Plume Model Validation and Development Project: Moderately Complex Terrain Site. EA-3755, the Electric Power Research Institute, Palo Alto, CA.

Bowne, N.E., 1987. Observations and Evaluations of Plume Models. TRC Environmental Consultants, Inc., East Hartford, CT.

Burton, C.J., Stoeckenius, T.E. and Nordin, J.P., 1982. Variability/ Uncertainty in Sulfur Emissions: A Summary of Current Knowledge on the Effects on Ambient Standard Attainment Demonstrations of Adopting Some Simple Models of Sulfur Variability. Proceeding of the Woods Hole Workshop on Modeling Uncertainty.

Carson, D.J., 1986. A Report on the Symposium on Uncertainty in Modelling Atmospheric Dispersion. Atmospheric Environment 20, 1047-1049.

Dupuis, L.R. and Lipfert, F.W., 1986. Estimating the Cost of Uncertainty in Air Quality Modeling. EA-4707, the Electric Power Research Institute, Palo Alto, CA.

Environmental Protection Agency, 1982a. Air Quality Criteria for Particulate Matter and Sulfur Oxides, Volume I. EPA600/8-82-029a, U.S. Environmental Protection Agency, Research Triangle Park, NC 27722.

Environmental Protection Agency, 1982b. Air Quality Criteria for Particulate Matter and Sulfur Oxides, Volume II. EPA-600/8-82-029b, U.S. Environmental Protection Agency, Research Triangle Park, NC 27722.

Environmental Protection Agency, 1986. Guideline on Air Quality Models (Revised). EPA-450/2-78-027R, U.S. Environmental Protection Agency.

Fox, D.G., 1981. Judging Air Quality Model Performance. Bulletin of the American Meteorological Society 62, 599-609.

Hanna, S.R., 1986. Spectra of Concentration Fluctuations: The Two Time Scales of a Meandering Plume. Atmospheric Environment 20, 1131-1137.

Hayes, S.R. and Moore, G.E., 1986. Air Quality Model Performance: A Comparative Analysis of 14 Model Evaluation Studies. Atmospheric Environment 10, 1897-1911.

Hillyer, M.J. and Burton, C.S., 1980. The ExEx Methods: Incorporating Variability in Sulfur Dioxide Emissions Into Power Plant Impact Assessment. Prepared for the U.S. Environmental Protection Agency, Research Triangle, Park, NC.

Hilst, G.R., 1978. Plume Model Validation. EA-917-SY, the Electric Power Research Institute, Palo Alto, CA.

Moore, G.E. and Liu, M.K., 1987. Time Series and Spectral Analyses of Concentration and Meteorological Data from the Kincaid and Bull Run Power Plants. SYSAPP-87/014, Prepared for the Electric Power Research Institute, Palo Alto, CA.

Murray, D.R. and Bowne, N.E., 1988. Urban Power Plant Plume Studies. EA-5468, the Electric Power Research Institute, Palo Alto, CA.

Smith, M.E., 1984. Review of the Attributes and Performance of Ten Rural Diffusion Models. Bulletin of the American Meteorological Society 65.

Taylor, J.A., Jakeman, A.J. and Simpson, R.W., 1986. Modeling Distributions of Air a Pollutant Concentrations --I. Identification of Statistical Models. Atmospheric Environment 20, 1781-1789.

Thrall, A.D., Stoeckenius, T.E. and Burton, C.S., 1985. A Method for Calculating Dispersion Modeling Uncertainty Applied to the Regulation of an Emission Source. Prepared for the U.S. Environmental Protection Agency, Research Triangle Park, NC.

Turner, D.B. and Irwin, J.S., 1982. Extreme Value Statistics Related to Performance of a Standard Air Quality

Simulation Model Using Data at Seven Power Plant Sites. Atmospheric Environment 16, 8. 1907-1914.

Venkatram, A., 1979. A Note on the Measurement and Modeling of Pollutant Concentrations Associated with Point Sources. Boundary-Layer Meteorology 17, 523-536.

Wilson, D.J., Robins, A.G. and Fackrell, J.E., 1985. Intermittency and Conditionally-Averaged Concentration Fluctuation Statistics in Plumes. Atmospheric Environment 19, 1053-1064.

MODELING POLLUTANT PLUMES

Robert W. Jernigan
The American University and
Statistical Policy Branch, EPA, Washington, DC

The previous papers and presentations by Anthony D. Thrall and Ralph E. Folsom show two different approaches to the modeling of pollutant plumes. Thrall is mainly concerned with the modeling of isolated air pollution point sources, such as coal fired power plants. As such the methodology that he considers involves models of the dynamic behavior of plume development and dispersion. These models are used to evaluate the effect of various control methods that are considered to meet the standards of the Clear Air Act Ammendments of 1970 and 1977. There is little to comment on in this paper. I have seen the details of some of Thrall's related work where he has been involved in developing statistical summaries of model performance and in the calibration problem of matching model calculations to observed monitoring data. Using a measure of the probability of misclassification derived from simulated data, he has developed quantitative procedures for the validation of such models.

Folsom gives us more to examine, so my comments will center on his paper. Folsom has been developing methodologies for the modeling of groundwater pollution plumes. His modeling efforts rely on a certain stability or stationarity in the pollutant plume. His procedures are used to aid in the design of well-site sampling plans. The use of indirect measurements of contamination obtained from electromagnetics and resistivity devices is the crucial element in his mapping of plumes through response surface methods. The use of such a covariate for the contaminant provides a useful and relatively inexpensive view of the pollutant plume.

Folsom uses a class of low order polynomial spatial regression models along with this covariate to predict pollutant concentrations. He does not however incorporate any measure of spatial autocorrelation in this estimation. Unfortunately, inclusion of a covariate does not necessarily explain all of the autocorrelation present. In fact one would see substantial reduction in spatial autocorrelation only if the relationship between the measurement and its covariate was not affected by spatial properties. For example it might happen that a measurement and its covariate are more highly correlated in areas of high rather than low pollution. An indication of such a behavior can be seen in the scatterplots of EM and resistivity measurements in Folsom, et.al.(1985) There a few high side outliers often behave quite differently from the bulk of the data.

Of course, the technique of co-kriging could incorporate both the information on the covariate and the spatial autocorrelation to model the contaminant plum. But it is easy to incorporate spatial autocorrelations and the essence of this kriging methodology into his response surface techniques without resorting to specialized kriging software. What is needed is a preliminary estimate of the spatial autocorrelations. These will be used in a generalized least squares algorithm to estimate the contaminant. This technique will yield predictions of measurements of the contaminant plume whose spatial autocorrelations can be calculated. These improved estimates of the spatial autocorrelations can then be used with generalized least squares to re-estimate the pollutant concentrations. As such an iterative technique continues it incorporates spatial effects directly and R^2 values in excess of the 63% obtained by Folsom could be possible.

Folsom's use of a random coefficient regression model is one approach to account for the within site sampling variability as well as the laboratory analysis variability. Kriging with a nugget effect to smooth the response surface is another way to model these sources of variability.

In all such modeling problems we often produce a map of point estimates. We must be careful in not interpreting this map as truth. Each point on our map may be above or below the true pollutant response surface. Our estimation error may be quite small and still the map of the true surface can appear different from that of our estimated surface, especially in the presence of spatial autocorrelation.

I applaud Folsom and his colleagues for considering extensions to his models to incorporate spatial autocorrelation. Perhaps by combining the existing techniques of mapping groundwater contaminants more efficient and reliable procedures can be developed to model pollutant plumes.

ESTIMATING THE SPATIAL UNCERTAINTY OF INFERRED RATES OF DRY ACIDIC DEPOSITION

Richard A. Livingston
Geology Dept.,U.of Maryland,College Park,MD 20742

The acid deposition monitoring program conducted by NAPAP concerns monitoring both wet and dry deposition. Wet deposition or acid rain monitoring has been in operation since 1982. The dry deposition monitoring program is now getting underway[1]. One of the major issues with monitoring dry deposition concerns the extrapolation of the dry deposition rate measured at a monitoring site to the entire region of interest.

The problem is further complicated by the fact that the measurement technique involved gives only an indirect estimate of the actual flux to the ground. Instead, an approach known as the inferential method is used. This consists of measuring concentrations of acidic gases and particles in the atmosphere. From the concentrations, the rate of deposition is inferred using an algorithm that incorporates several site-specific or time-dependent factors[2]. For example, the type of ground cover, which varies spatially, is a factor in the algorithm. An example of a time-dependent factor would be windspeed. These factors are measured at the actual monitoring site, but can obviously assume different values as one moves away from the site.

The inferential method employs a simple formula for calculating dry deposition rates:

$$F = V_d C \qquad (1)$$

where F is the dry deposition flux; C is the concentration of the acidic species in the atmosphere, and V_d is a factor known as the deposition velocity which is calculated from the algorithm described above. The deposition velocity can range from 0.01 to 1 cm/s while the concentration of a gas such as sulfur dioxide is in the range of 1 to 50 $\mu g/m^3$.

The difficulty is that both terms on the righthand side of Equation 1 can vary spatially. Consequently, the dry deposition flux, F, can also vary spatially. However, the problem can be posed in such a way that effectively decouples the variability in C from the variability in V_d.

Differentiating equation 1 with respect to some spatial coordinate, x, gives:

$$\frac{dF}{dx} = V_d \frac{dC}{dx} + C \frac{d(V_d)}{dx} \qquad (2)$$

It is generally the case that at least in rural areas far from major point sources or urban areas, the gradient of the atmospheric pollutant concentration is very shallow, on the order of 10 $\mu g/m^3$ on a scale of kilometers. On the other hand, for reasons discussed below, the deposition velocity can change by an order of magnitude over a distance of a meter or less.

Thus, we can make the assumption that:

$$V_d \frac{dC}{dx} << C \frac{d(V_d)}{dx} \qquad (3)$$

and hence:

$$\frac{dF}{dx} \approx C \frac{d(V_d)}{dx} \qquad (4)$$

In other words, on a local scale, the spatial variability of the dry deposition is determined primarily by the variations in the deposition velocity since the air pollution concentration can be regarded as constant.

While this reduces the complexity of the problem, it also creates some awkwardness since it implies that the major variability is not with the variable that is being monitored. Thus it is not possible to estimate the spatial uncertainty of dry deposition simply from the data on atmospheric concentrations. Instead one must estimate the variability of the deposition velocity, which entails acquiring additional information.

It is now necessary to consider in more detail the factors affecting the deposition velocity. As noted above, type of ground cover is an important factor. The deposition velocity over bare rock will be different than that over vegetation. In fact, there are further subdivisions under each category. For example, a bare limestone will react more readily with acidic gases than bare granite.

With vegetative ground cover, there are additional considerations. The leaves of different species of plants have different types of surface chemistry and pores, or stomata, for uptake of acidic gases. Thus, the deposition velocity on a pine tree will differ from that on an oak tree. Aside from the surface chemistry, the physical structure of the plant is also important in determining the deposition velocity. The spatial pattern of the leaves and branches affects the wind velocities and hence the deposition velocities in the forest canopy. For the same reason, grasslands have a different deposition velocity than forested areas.

Not only is the type of ground cover important, but also the spatial relationships of one type of ground cover to another. An abrupt change in the height of ground cover such as occurs going from a grassy field to a wooded area will create additional turbulence in air flow patterns with a resulting higher deposition velocity. Thus, a uniform groundcover will have a deposition velocity different from that of an area where patches of one type of ground cover are intermingled with another.

Also affecting the wind field and thus the deposition velocity is the topography. A hilly area will have a much greater variance in windspeeds and directions than a flat area. The orientation to prevailing winds and to solar radiation can influence deposition velocities as well because deposition velocities are influenced by surface temperature and wetness.

Thus the problem of characterizing the spatial variability of the dry deposition comes down to being able to estimate the spatial variability of all these factors that determine the deposition velocity. To do this, it is neces-

sary to assemble data and to use statistical methods from widely diverse disciplines.

The mapping of vegetation types is a long established part of plant ecological studies. A variety of data bases already exist at Federal, state or local levels on land use and on the types of vegetation, at least for selected areas. One example is the Geoecology data base established at Oak Ridge National Laboratory. It remains to be seen whether these existing data bases have fine enough spatial resolution to be applied here. There is also the question of whether the sampling procedures were consistent enough among the different data bases to permit pooling of the data. The possibility also exists of creating new data bases through interpretation of aerial photography.

Ecologists have developed a number of statistical techniques for dealing with spatial patterns of plant species. These include indices of species diversity, species abundance and patchiness[3]. As yet, these measures have not been applied to the specific problem of dry deposition.

Similarly, there is an abundance of topographical maps on different scales. However, these generally do not exist in digitized form. Geomorphologists have developed their own set of statistical measures to characterize the physical landscape. Some relevant indices of drainage basin morphometry include drainage density, relative relief and ruggedness number[4].

In addition to making use of these existing statistical tools, it will probably be necessary to develop new methods. For example, fractal analysis may prove very useful for characterizing certain aspects of the landscape. For instance, a forest canopy has many of the attributes of a fractal surface.

Integrating all this information will no doubt require the use of computers. Specialized methods for dealing with spatial data have been developed under the general heading of Geographic Information Systems. The rapidly growing field of pattern recognition may also provide some assistance.

Ultimately, once all the uncertainties in the individual factors have been estimated, it will then be necessary to make some summary statements about the overall spatial uncertainty of the inferred dry deposition rates. In essence, the process outlined here consists of taking the individual uncertainties in the underlying factors and propagating them through the algorithm that calculates dry deposition. It should be noted that this process yields only an estimate of the precision of the inferred dry deposition. The bias is impossible to determine absolutely since there is no independent measurement technique that could be used as a reference method. However, it may be possible to put some bounds on the bias through the use of mass balance estimates over a controlled watershed or over a specific area of ground cover.

Although it is not possible to state an absolute value for the bias and hence the true rate of dry deposition, the data can from the dry deposition network will still be useful for policy-making. That is because the critical policy questions being asked tend to be of a relative nature, such as: How much greater is the dry deposition rate over affected areas compared to unpolluted background regions. Or: Is dry deposition getting greater or lesser over time. These questions thus often involve differences in time or space rather than absolute values. If the bias is constant, it will be removed in calculating the differences.

At this point another statistical issue must be dealt with. Given the variability in the factors underlying the deposition velocity, it is likely that the estimated precision will also vary spatially. The characterization of variation over a region is a classic problem in geostatistics, and a number of methods have been developed[5]. The most well known method is kriging.

However simple kriging is not useful for this purpose since it provides no insight on the actual uncertainty at any distance from the actual measurement points. More advanced techniques such as indicator or probability are now coming into use. These methods can generate maps that show areas where some minimum level of detection is being exceeded[6].

In conclusion, the task of estimating the spatial uncertainty of dry deposition depends heavily on the use of auxiliary information. This information comes from a variety of disciplines, each with its own statistical tools. Integrating the information from these diverse sources and presenting it in a meaningful way for policymakers will be a major challenge for the statistical community.

References

1. National Acid Precipitation Assessment Program (1987): Annual Report 1986, US Government Printing Office, Washington DC, pp. 75-88
2. Hicks,B.,Baldocchi,D.,Hosker,R., Hutchison,B.,Matt,D.,Mcmillen,R., and Satterfield,L.(1985):"On the Use of Monitored Air Concentrations to Infer Dry Deposition", NOAA Technical Manual ERL ARL-141, Air Resources Laboratory, Silver Spring, MD, 65 pgs.
3. Pielou,E.,(1975): Ecological Diversity, Wiley-Interscience, New York, 165 pgs.
4. Ritter,D.(1978): Process Geomorphology, Wm.C. Brown Pubs., Dubuque Iowa, pp. 154-176.
5. Journel,A. and Huijbregts,Ch.,(1978): Mining Geostatistics, Academic Press, London, pp. 303-444.
6. Journel,A.(1984): Indicator Approach to Toxic Chemical Sites, US EPA Environmental Monitoring Systems Laboratory, Las Vegas, NV, 30 pgs.

ASSESSING EFFECTS ON FLUCTUATING POPULATIONS: TESTS AND DIAGNOSTICS

Allan Stewart-Oaten
Biology Dept., University of California, Santa Barbara, Ca., 93106

1. Introduction

A standard question in environmental impact assessment is whether some human alteration has caused a decrease in some environmental variable.

There are some obvious difficulties. An apparent change over time could be due to something else occurring at about the same time as the alteration, or to a general trend over time. A lower value of the variable near the alteration than far away could be due to other differences between the locations.

One way to deal with these problems is to have observations both Before and After the alteration, at sites both near to it (Impact) and far away (Control). (These capitalized words will refer to these fixed periods and sites henceforth.) The plan is to compare the difference between Impact and Control sites Before with the difference After, and to argue that a change in this difference cannot be the result of either a widespread time change or of a difference between the sites, so is likely to indicate an effect of the alteration.

There are some obvious difficulties with this scheme too. One is that a change in the difference could be due to some time-related change other than the alteration, but occurring at about the same time and affecting one site but not the other. Another is that the difference may fail to change because the alteration affects both sites about equally. Thus the Control needs to be close enough to the Impact site to be affected similarly by natural changes, but far enough away to be little affected by the alteration. This important task is not the subject of this paper, and we will assume that it has been accomplished - while recognizing that this assumption needs to be examined and can never be guaranteed.

A more technical problem is that many environmental variables fluctuate naturally over time. This applies to physical or chemical variables, but perhaps especially to the densities of biological populations, which will be our focus. Thus any observed difference between sites can be due to any of three causes - sampling error, natural fluctuations, and intrinsic differences between the sites - of which we wish to distinguish the third. This cannot be done with only one Before sampling time (or only one After, unless we make the risky assumption that variability due to the other causes is the same After as Before), so we assume there are many sampling times in each period.

It is also difficult to do if the two sites are sampled at different times. One or other site is likely to have been sampled at times of generally greater abundance, so that one must either include covariates (frequently having little idea which ones matter or how to model them) or pretend that the sampling times were chosen and assigned randomly and then the records of them were lost. We will therefore assume that the two sites are sampled simultaneously, or nearly so.

We will thus have observations X_{ijkm}, for i = 1 (Before) or 2 (After), j = 1, 2, ..., n_i (sampling times in period i), k = 1 (Impact) or 2 (Control), and m = 1, ..., r_{ijk} (replicates).

In this paper, we will assume r_{ijk}=1. The replicates give information on sampling variance but not on fluctuation over time. If the r_{ijk}'s are equal and the sampling errors are iid Normal, with known variance, the replicates can be replaced by their average (a sufficient statistic) without loss of information. The same is true if the sampling variance is unknown, but the time fluctuations (or their differences between sites) are also iid Normal. Neither of these conditions will usually hold, so the replicates might show us a better "summary" than the average, and might contribute information useful for model building or for estimating variances in which both sampling error and time fluctuations play a part. However, this approach seems likely to be quite complex, and may not be worth the trouble if the time fluctuations are large.

Environmental impact assessment is often cast as a problem in statistical testing: the task is described as one of detecting a change or a violation of some threshhold, and one of the most frequent techniques invoked is Analysis of Variance, which is primarily a testing (rather than an estimation) procedure. Section 2 below discusses the testing problem, particularly the assumptions for a standard two-sample test on the Impact-Control differences, and ways to adjust the data to satisfy them, if necessary. Some data will be presented to illustrate the effectiveness of these ways, and some of their shortcomings, in Section 3. Section 4 briefly considers three topics: gaps in the present testing scheme and directions in which it needs to be extended; weaknesses of testing as a goal, and an approach to the more appropriate goal of estimation; and some implications for the design of assessment programs.

2. Testing

A natural model for the observations, X_{ijk}, is

$$X_{ijk}=m_i+a_{ij}+b_k+c_{ik}+e_{ijk} \qquad (1.1)$$

where m_i is the overall mean for the period, a_{ij} is a time effect, b_k is a site effect, c_{ik} is c if i=2 and k=1 but is zero otherwise, and gives the change in the site effect between periods, presumably due to the alteration, and the e_{ijk}'s are iid errors. If we believed this model, and wanted to estimate c, we would use two-sample test statistics derived from the differences, $D_{ij}=X_{ij1}-X_{ij2}$.

This procedure remains valid under weaker assumptions than those in (1.1). These assumptions cannot be guaranteed but can be checked for plausibility against both biological theory and the observations themselves. They are:

(1) The D_{1j}'s all have the same mean, M_1. That is, model (1.1) does not need a time-x-site interaction term, d_{ijk}, so $M_1=b_1-b_2$. If there is no effect of the alteration, the D_{2j}'s also all have this mean. If the alteration does have an

effect, the model says that the D_{2j}'s all have mean $M_2=b_1-b_2+c$, but the test would not be invalidated if c varied over time, though it may well be inefficient.

(2) The errors, $E_{ij}=e_{ij1}-e_{ij2}$, are independent.

(3) The E_{1j}'s are identically distributed; in particular, they have equal variances. The E_{2j}'s also are identically distributed.

(4) The distribution of the E_{1j}'s is the same as that of the E_{2j}'s.

(5) The E_{ij}'s are Normal.

These assumptions are not all essential, or equally important.

The first seems the most important. If it does not hold, at least approximately, then neither "the" Before nor "the" After mean difference, whose comparison is the aim of the test, is well-defined.

In one sense, this assumption can be guaranteed. The errors, e_{ijk}, will be the result of both sampling error and the "random" fluctuations over time of the actual populations. The distinction between "fixed" and "random" time effects seems necessarily arbitrary. Individual births and deaths, and probably movements, of individuals or schools, seem best regarded as random. But the rates of these changes may be altered by storms, upwellings or other physical or chemical fluctuations. These may be to some extent predictable and can sometimes be long-lasting, which suggests fixed effects, but they are irregular and hard to predict, so can be argued to be random. Finally, all these sources of change may depend on seasons, which seem predictable enough to be called fixed - except that their effects on the organisms seem to vary in timing and amount from year to year.

It may thus be possible to argue that time effects should all be classified as "random", in which case they and time-x-site interactions can all be regarded as part of the error. Assumption (1) is then formally justified. But this is only a delaying tactic: any test results can then be attributed to the effects of the particular sampling times chosen, and the problem must be dealt with again. It also reduces the plausibility of assumption (2), since major events like seasons or long-lasting weather conditions (e.g. El Nino) will affect several sampling times.

Another approach might be to try to list the time-related effects to be called "fixed", and explicitly allow for them in the model. But we usually do not know which effects to include or how to model them. The response of many species to even the most obvious candidate, seasons, is not consistent, especially in timing, and the shape of the response can rarely be modelled by a simple function like a sine wave.

These problems vanish if large or long-lasting time-related effects are approximately the same at the two sites, so they cancel in the differences: that is, if these time-related effects are approximately additive with respect to site.

The idea is that there is a "region", containing Impact and Control, over which major time-related changes, such as seasons or major weather conditions, are imposed approximately uniformly. The problem is that these uniform changes are imposed on locations that differ, so they may not have uniform effects. To these effects are added the fixed effects of the locations themselves, and local fluctuations: local time-related changes such as minor weather effects or water movements. Together these comprise the external forces influencing the path of the population density over time. The actual population is determined by "random" events like individual deaths, births or migratory movements, but these stochastic events occur at rates governed by the external forces together with "internal" forces resulting from the size of the population itself, and of other interacting populations, in the recent past.

Treating the uniform, major changes as given, the path traced out by the actual population density over time at a given site can be seen as a stochastic process whose mean function is determined by the uniform changes, and whose deviations from this function are due to the local or internal forces, or to the chance events whose probabilities are determined by all these forces. The assumption that the Impact-Control differences have equal means over time requires that the mean function at Impact differ from that at Control by a constant amount: the differences between the two sites should not lead to differences in the effects that the major time-related changes have at these sites.

It is the mean functions that need to be additive in this sense, not the actual populations themselves, which incorporate the deviations as well. For example, additivity cannot be tested by taking replicate samples at each time and place, and comparing the variation among replicates with the variation due "interaction" between the two sites and the n_j sampling times. This tests non-additivity in the actual population values (which we expect because of the deviations) but not in the mean functions. Even if we knew the actual populations at each time, we would still need to look for non-additivity in a 2-way table (times and sites) with one observation per cell. The addition of sampling error does not change this basic situation: the only difference is that we must use estimates of the actual population instead of known values.

With only one observation per cell, we cannot test for all possible interactions, so we choose the most likely. This may depend on the situation and the species, but in most cases we would expect non-additivity to show up as a relation between the difference between the two sites and the overall abundance of the species. An obvious measure of overall abundance is the average of the two sites, so we are led to look for a monotone relation between the difference and the average. Two simple test statistics for this are Spearman's rank correlation coefficient and the slope of the regression of the differences against the averages - which corresponds to Tukey's (1949) non-additivity test for this situation.

As these tests imply, we would expect time-related changes to have larger effects at dense sites than at sparse ones, for many species. Many biologists would expect these changes to be

multiplicative, and would suggest using the logs of the sample abundances. Though appealing, this suggestion may be the result of misplaced intuition (Stewart-Oaten 1986). Additive changes in rates will lead to multiplicative changes in density for exponentially increasing or declining populations, but there is no obvious guarantee that the changes in rates are additive and most biological populations are less likely to be changing exponentially than to be at a rough equilibrium, fluctuating in response to changes in carrying capacity.

Rather than pre-judge this issue, a range of transformations, including logs, can be considered. Probably the best-known and most convenient family is the one used by Box and Cox (1964),

$$Y(X)=((X+c)^b-1)/b \qquad (1.2)$$

which can be thought of as the shift-plus-power transformation, $(X+c)^b$ when b=0 and $\log(X+c)$ when b=0. In this paper, b is restricted to lie between -1 and 1. Larger values of b imply that a given time-related input will change the untransformed abundance more at the sparser site than at the denser. Smaller values imply that the amount of habitat (volume of water or area of land) required to support a single animal will change more at the site where it is smaller. Both seem unlikely.

Assuming there are values of b and c for which Y is additive, homoscedastic and Normal, Box and Cox proposed estimating them by maximum likelihood or Bayesian methods. There are some difficulties with this approach. A technical one is that, at least for some values of b, the likelihood is not a well-behaved function of c (Berry 1987). A conceptual one, first raised by Tukey's discussion of Box and Cox (1964), is that the method may emphasize Normality, homoscedasticity and additivity in that order (especially when there is only one observation per cell), when the order of importance is generally agreed to be the reverse. It is often asserted that these properties tend to go together (e.g. Atkinson 1985, p. 81), but this claim seems to rest on a small set of non-randomly chosen examples.

In this paper, we propose first assessing additivity by the Tukey and Spearman tests, and choosing a transformation from among those for which these tests (and accompanying plots) do not indicate significant non-additivity in the Before period. (Non-additivity After might be an effect of the alteration. However we may want to consider it if the After observations are different overall: for example, if the Control area has clearly changed, implying the whole region has, for some reason other than the alteration. Our transformation choice method is very like fitting a regression, and we face the same risks when we extend the estimated relationship well beyond the observed range.)

The null hypothesis for these tests is really stronger than additivity: the differences and averages will also be correlated if the Impact and Control means are identical but the variances differ. However, for population counts, this is likely to lead us astray only if the variation in the averages, due to uniform time-related effects, is small compared to any difference between Impact and Control in variation due to local time-related effects and sampling error. For most species, this seems far from the case.

In practice, we are likely to find that either many transformations seem approximately additive or none do. The latter case could arise if Impact is denser than Control under some conditions (e.g. when overall abundance is high) and sparser under others. For such a species, the alteration may also be beneficial under some conditions but not others, and it may be best to carry out separate analyses for the two sets of conditions. The possibility of such a time-x-site interaction that cannot be transformed away forms part of the motivation for methods like those of Section 4.

If many transformations are approximately additive, we must choose on other grounds. Equality of variances - the main aspect of our assumption (3) - is widely agreed to be the next most important goal, and there are reasons for thinking this to be the case here.

Often this assumption is not crucial. If we have additivity and independence, we have observations Y_{ij} with distributions F_{ij}, having means m_i. It is natural to compare this with the homogeneous situation where Y_{ij} has distribution $G_i = n_i^{-1} \Sigma F_{ij}$. Stigler (1976) shows that any linear function of the order statistics has the same mean and at least as small a variance in the heterogeneous situation as in the homogeneous one. These results include most location estimates and some scale estimates. They also hold for the standard variance estimate. Thus a t-like test, with a measure of discrepancy in the numerator and an estimate of its standard deviation in the denominator, should have approximately the same mean but a smaller variance than in the homogeneous case. Type 1 error probabilities will then be decreased and Type 2 error probabilities will be increased for alternatives near the null but decreased for distant alternatives.

This comparison with the homogeneous case ignores two points. One is that another transformation, for which the corresponding F_{ij}'s are more alike, may be more efficient - may have more power to detect effects of the alteration. This seems impossible to decide in general. If there is an effect, then Impact and Control are different either Before or After or both, so at least one of the transformations is non-additive in at least one period. The comparison would also depend on which tests were used, and on efforts to allow for heterogeneity by weighting the observations (Box and Hill 1974).

Perhaps the main problem is that the alteration may have different effects at different times. For example, an alteration that warms part of a body of water might be producing a temperature that the animals prefer in Winter but dislike in Summer. If Impact and Control are approximately equally abundant in the Before period, differences and averages will appear unrelated for almost any transformation. But it still may be that the size of a time-related effect depends on the size of the local population it is operating on. If populations are smaller in Winter, when the alteration has a positive (or large) effect, than in Summer, when the effect is negative (or small), the results of

tests and estimates may depend strongly on the transformation choice.

This is not a simple issue. In a case like this, the results will also depend on which season is better represented in the sampling times chosen. Also, it is not clear how "the" effect should be assessed even if we know the pattern: does the positive effect in Winter, when the local population's survival or genetic structure may be threatened, outweigh the negative effect in Summer, when the population's interactions with its environment are at their peak?

A workable, though imperfect, solution may be to carry out three analyses, one for all the data, another for Summer (or "abundant" periods as determined by, say, the Control value), and the third for Winter. For each analysis, the choice of transformation (among those that appear to be additive) could be determined partly by whether absolute deviations from the (supposedly) common mean tend to increase, decrease or stay the same as overall abundance (measured by the average of Impact and Control) increases.

Different variances for the Before and After differences (assumption 4) do not seem to affect the validity of t-tests much (Posten 1982), if the sample sizes are nearly equal. If they are not, the Welch-Satterthwaite modification seems almost always to give a conservative test. Non-Normality also seems to have little effect on the validity of the t-test (e.g. Posten 1978). The t-test is not efficient for many non-Normal distributions, but this problem seems best approached by using more robust tests along with it.

The results on which these comments are based are for symmetric situations: either the errors have symmetric distributions, or the Before and After errors have the same distribution (and sample size). One-sided one-sample t-tests involving skewed distributions may have actual levels far from the nominal ones (Posten 1979), though this seems to occur only for small samples or extreme skewness. This suggests, though, that two-sample tests might go astray if the Before and After difference distributions are skewed and not the same shape. In such a case, also, the various robust tests we might use in concert with the t-test - e.g. the Wilcoxon-Mann-Whitney test, or the t-like test based on the 25% trimmed mean, which will be used here - will be testing different things: the true Before and After means, the true trimmed means, and the true probability that a random Before difference will exceed a random After difference may not all agree on whether Before differences are "larger" than After differences.

Tests and plots for symmetry and for comparing the shapes of the distributions of the Before and After differences will thus be useful, partly for deciding whether a particular transformation is "acceptable" but more for indicating what to do if tests for different acceptable transformations disagree. There are many tests for symmetry, most of which are too prone to reject symmetry for symmetric distributions with long tails. Gupta's (1967) test does not seem to do this (Antille, Kersting and Zucchini 1982). A very crude measure of agreement between the shapes of the distributions of the Before and After differences (independent of differences of location or scale) is obtained by standardizing each observation (subtracting the mean and dividing by the standard deviation) and calculating the standardized Smirnov statistic, $(n_1n_2/(n_1+n_2))^{1/2}D$, where D is the maximum difference between the cdf's (Lehmann 1975, p. 37). If this measure is under about 1, the shapes seem compatible (Lehmann, Table F). (Because the standard deviation may be a poor measure of scale for some distributions, I have removed the top and bottom 10% of each distribution before calculating this scale measurement in Section 3's examples. A criterion smaller than 1 might be desirable after all this matching.)

Assumption (2) remains. The differences form a time series, and would be expected to be serially correlated if the sampling times were very close together. On the other hand, if we have an additive transformation, the major time-related effects may cancel when we take the Impact-Control differences, and the minor, local effects remaining may not be long-lasting. (A more detailed argument of this kind appears in Stewart-Oaten, Murdoch and Parker 1986 and in Stewart-Oaten 1986.) There are many ways to test for serial correlation. Two simple ones are the von Neumann ratio (von Neumann 1941; Anderson 1970, Table 6.3), the predecessor of the Durbin-Watson test, and the rank version of this test (Bartels 1982).

If these tests indicate that serial correlation has not been removed, we have a difficult problem: a time series whose observations are not (usually) equally spaced and, whether they are or not, is likely to be non-stationary because the magnitude and frequency of disturbances, and the vulnerability of the population, vary over time. The serious modelling needed is beyond the scope of this paper. In an example below (Species 3), I have simply adjusted the observations to approximate independence in a very ad hoc manner. (For cases where the series is amenable to ARIMA modelling or to ARMA modelling with a fixed seasonal component, see Box and Tiao 1965 and 1975, and Reinsel and Tiao 1987.)

Thus a scheme for testing for an effect of the alteration is: transform the raw X_{ijk}'s to additive Y_{ijk}'s as indicated by the Tukey and Spearman tests and check for a relation between the averages ($Y_{ij.}$) and the absolute deviations of the differences ($Y_{ij1}-Y_{ij2}$) from their median. In the examples below, this relation was examined by computing the p-values for the slope of the regression of the absolute deviations against the averages and the Spearman test for correlation between them. Check the differences for serial correlation, and allow for it if present. Carry out a two-sample test to see if the differences have become smaller following the alteration. Check whether the test result is likely to be affected by skewness, using Gupta's test and a measure of the discrepancy between the shapes of the differences of the transformed Before and After values. Also check whether the result may be due to changes other than the alteration; one check is to regress the Before differences against time

Because the t-test lacks robustness, it makes sense to use robust tests as well. Similarly, it

makes sense to use a range of apparently acceptable transformations. If the results disagree, we need further thought as to which ones most nearly describe the changes of concern. If the results agree, we need to think about the size of the effect and mechanisms. This will be easier if the transformed variables are in interpretable units: b=1 (change in absolute numbers), b=0 (percentage change) and b=-1 (change in the amount of space needed per animal) seem best.

3. Examples

The following examples all use real data from an ongoing assessment project. However, I have rescaled all numbers so that the maximum in each data set is 95 and have interchanged "Impact" and "Control" in some cases.

Three kinds of table appear. Data tables list Impact ("I") and Control values by sampling time (read down the columns first). Times were not listed unless the independence checks indicated a serial correlation problem or there were separate analyses for different "seasons". They are given in units of a year, beginning with January 1 of the first year the species was sampled.

Preliminary checks give the power and constant of the transformation, the p-values for the Tukey and Spearman tests (differences vs. averages), the slope and Spearman tests for a relationship between the average and the absolute deviation of the difference from the median difference, the von Neumann test for serial correlation and its rank version, Gupta's test for symmetry, and the test for a trend with time. "Shape" is the comparison of the shapes of the distributions of the Before and After differences. All tests are two-sided except those for serial correlation. Negative "p-values" indicate a negative slope or correlation of the tested variable against the averages. All transformations are monotone increasing: when b is negative, the transformation is $-(X+c)^b$.

Four tests for an effect of the alteration were used: the standard t-test, the Wilcoxon-Mann-Whitney test, and the t-like tests based on the 25% trimmed mean (Yuen and Dixon 1973) and on the biweight estimate (Kafadar 1982; this paper is missing the factors $[9\ MAD]^2$ and $[6\ s_{bi}]^2$ in the definitions of s_{bi} and S_T respectively; I have used 4 instead of 6 in her equ (2), see Kafadar 1983).

When a non-positive power was used, and the constant was zero, all sampling times with a zero at either Impact or Control were omitted from the analysis. Sampling times with zeros at both sites were omitted from all analyses, though they are listed with the data.

In all plots, filled squares indicate Before data and empty squares After data.

The data and analyses presented here do not support any conclusions relevant to the actual project. First, as already stated, Impact and Control have sometimes been reversed. Second, the analyses are all purely numerical. I have made no use of anything known about the species involved, or even given their names. In practice these analyses should be used in concert with information about the physiology and population dynamics of the species concerned. Third, the examples here have been selected from data on hundreds of species, to illustrate statistical problems: they are not typical.

TABLE 1: Data for Species 1.

BEFORE

I	C	I	C	I	C
0.83	0.68	0.25	0.52	28.62	34.75
2.17	0.87	1.09	0.80	1.35	2.60
0.11	0.39	0.98	2.39	0.15	0.16
0.37	0.19	14.83	10.00	0.41	0.09
3.13	5.22	1.42	0.62	1.46	1.31
0.10	0.14	0.07	0.15	1.19	0.71
0.08	0.03	0.16	0.07	1.55	1.18
0.73	0.44	0.49	0.21	3.19	5.27
15.14	1.31	0.00	0.02	4.77	1.15
4.68	2.46	5.56	3.79		
0.00	0.00	41.57	12.43		

AFTER

I	C	I	C	I	C
1.84	0.87	0.34	0.02	2.45	4.57
0.32	0.64	0.00	0.00	5.41	5.69
0.54	2.14	1.34	22.29	75.26	35.81
0.02	0.16	0.11	0.37	95.00	27.95
0.08	0.47	3.54	11.18	1.18	0.04
0.81	8.45	64.91	10.59	0.02	0.05
0.13	0.34	0.79	1.92	5.38	3.82
0.74	1.20	0.01	0.01		

Species 1. The data are in Table 1. The Before samples are spread over 5 years, and the After over 3. Table 2 shows that serial correlation is not a serious problem. However, no

TABLE 2: Preliminary checks for Species 1, all times.

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	0.00	0.00	0.28	0.00	0.00	0.88	0.89	0.91	0.32	0.82
-1.00	0.50	0.73	0.85	0.15	0.13	0.89	0.85	0.40	0.48	0.69
-1.00	1.00	0.58	0.81	0.82	0.93	0.71	0.85	0.71	0.42	0.89
-1.00	2.00	0.42	0.46	0.20	0.19	0.47	0.72	0.53	0.45	0.95
-0.50	0.20	0.63	0.89	0.14	0.14	0.88	0.90	0.31	0.41	0.73
-0.50	0.40	0.51	0.79	0.96	0.94	0.75	0.90	0.58	0.17	0.84
-0.50	0.50	0.47	0.75	0.72	0.76	0.69	0.85	0.71	0.23	0.89
-0.50	1.00	0.34	0.35	0.19	0.15	0.50	0.76	0.58	0.21	0.98
0.00	0.00	0.58	0.88	0.75	0.61	0.74	0.86	0.68	0.20	0.83
0.00	0.05	0.26	0.48	0.60	0.75	0.70	0.86	0.76	0.15	0.90
0.00	0.10	0.24	0.43	0.27	0.26	0.64	0.85	0.71	0.13	0.91
0.00	0.20	0.20	0.28	0.09	0.05	0.56	0.82	0.71	0.27	0.96
0.00	0.40	0.16	0.24	0.03	0.01	0.48	0.72	0.71	0.31	0.98
0.00	1.00	0.10	0.14	0.00	0.00	0.41	0.63	0.53	0.47	0.92
1.00	0.00	0.01	0.12	0.00	0.00	0.80	0.64	0.89	0.23	0.87

TABLE 4: Preliminary checks for Species 1, abundant times.

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	0.00	0.63	0.30	0.03	0.08	0.49	0.73	1.01	0.50	0.58
-1.00	0.50	0.77	0.49	0.09	0.09	0.58	0.69	0.62	0.39	0.58
-1.00	1.00	0.88	0.60	0.17	0.15	0.63	0.71	0.54	0.44	0.57
-1.00	2.00	0.95	1.00	0.34	0.62	0.68	0.66	0.54	0.44	0.56
-0.50	0.20	0.98	0.73	0.24	0.20	0.66	0.70	0.54	0.50	0.56
-0.50	0.40	0.97	0.76	0.30	0.28	0.67	0.70	0.54	0.50	0.56
-0.50	0.50	0.94	0.92	0.33	0.47	0.68	0.66	0.54	0.50	0.56
-0.50	1.00	0.84	0.94	0.49	0.65	0.70	0.66	0.35	0.44	0.55
0.00	0.00	0.64	0.50	0.92	0.95	0.74	0.73	0.43	0.50	0.56
0.00	0.05	0.63	0.50	0.95	0.95	0.74	0.73	0.35	0.50	0.56
0.00	0.10	0.62	0.45	0.98	0.91	0.74	0.73	0.35	0.50	0.56
0.00	0.20	0.60	0.39	0.96	0.76	0.75	0.73	0.35	0.50	0.56
0.00	0.40	0.56	0.39	0.86	0.50	0.75	0.73	0.35	0.50	0.56
0.00	1.00	0.48	0.33	0.63	0.39	0.76	0.75	0.35	0.50	0.57
1.00	0.00	0.16	0.45	0.01	0.00	0.85	0.83	0.70	0.33	0.87

positive power gives samples that appear either additive or homoscedastic. (I omitted from the table many of the transformations that failed these tests. The results for positive powers given here are typical of the others.) Once we restrict to the transformations that seem acceptable by the preliminary checks, the two-sample tests all indicate a significant change (Table 3).

TABLE 3: Species 1 tests, all times.

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	0.00	0.400	0.012	0.187	0.005
-1.00	0.50	0.037	0.003	0.012	0.001
-1.00	1.00	0.015	0.005	0.011	0.007
-1.00	2.00	0.011	0.015	0.024	0.029
-0.50	0.20	0.041	0.004	0.014	0.003
-0.50	0.40	0.022	0.008	0.010	0.010
-0.50	0.50	0.019	0.009	0.013	0.014
-0.50	1.00	0.017	0.023	0.034	0.041
0.00	0.00	0.061	0.019	0.031	0.037
0.00	0.05	0.055	0.028	0.043	0.059
0.00	0.10	0.051	0.027	0.054	0.071
0.00	0.20	0.051	0.033	0.065	0.092
0.00	0.40	0.058	0.050	0.070	0.124
0.00	1.00	0.083	0.052	0.094	0.197
1.00	0.00	0.806	0.065	0.215	0.174

While this seems convincing, we may want to look deeper because the largest Impact-Control differences actually occur in the After period. While these are at times of high overall abundance, we might wonder whether effects at "abundant" times are different from those at "sparse" times. Accordingly, I divided the sampling times into those for which the Control was above its median of .867, and those below. (Table 1's values are rounded. The second sampling time in the Before period was "abundant", while the first After time was "sparse".)

Table 4 shows that the transformations that appear acceptable overall do so for abundant times too, with similar test results (Table 5). (The trim-t values might be affected by the small samples.) Table 6, however, shows few transformations as acceptable on all counts for the sparse times: some promising ones give mildly skewed Before distributions. Test p-values are larger than those conventionally required for "significance" (Table 7), though they also suggest a reduction due to the alteration. The t-test here is affected by the long upper tail of the After differences.

TABLE 5: Species 1 tests, abundant times.

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	0.00	0.005	0.004	0.113	0.016
-1.00	0.50	0.005	0.006	0.129	0.018
-1.00	1.00	0.006	0.007	0.124	0.021
-1.00	2.00	0.009	0.016	0.096	0.028
-0.50	0.20	0.008	0.014	0.136	0.027
-0.50	0.40	0.010	0.016	0.124	0.030
-0.50	0.50	0.011	0.016	0.119	0.032
-0.50	1.00	0.015	0.030	0.104	0.041
0.00	0.00	0.047	0.052	0.135	0.083
0.00	0.05	0.049	0.058	0.136	0.086
0.00	0.10	0.051	0.058	0.137	0.089
0.00	0.20	0.056	0.058	0.139	0.095
0.00	0.40	0.064	0.064	0.154	0.107
0.00	1.00	0.089	0.064	0.192	0.138
1.00	0.00	0.789	0.216	0.451	0.047

TABLE 6: Preliminary checks for Species 1, sparse times.

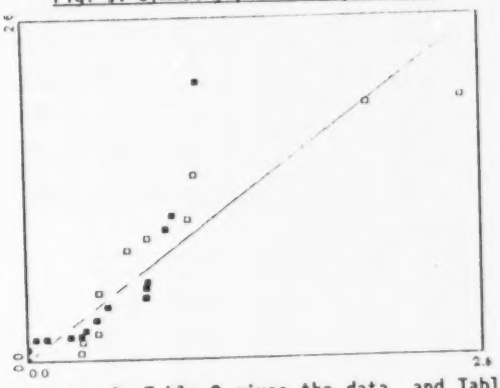
power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	0.00	0.01	0.64	0.00	0.00	0.85	0.94	0.43	0.29	0.97
-1.00	0.50	0.28	0.31	0.39	0.32	0.87	0.92	0.60	0.10	0.29
-1.00	1.00	0.11	0.07	0.79	0.93	0.84	0.80	0.60	0.23	0.28
-1.00	2.00	0.03	0.05	0.62	0.49	0.79	0.77	0.44	0.43	0.29
-0.50	0.20	0.35	0.44	0.24	0.15	0.87	0.92	0.64	0.05	0.31
-0.50	0.40	0.19	0.15	0.52	0.70	0.86	0.86	0.60	0.14	0.28
-0.50	0.50	0.15	0.13	0.63	0.82	0.85	0.79	0.60	0.20	0.28
-0.50	1.00	0.05	0.05	0.84	0.56	0.81	0.77	0.56	0.31	0.29
0.00	0.00	0.79	0.74	0.09	0.04	0.88	0.85	0.71	0.09	0.31
0.00	0.05	0.26	0.42	0.18	0.13	0.87	0.87	0.60	0.03	0.35
0.00	0.10	0.21	0.21	0.35	0.48	0.86	0.85	0.73	0.04	0.31
0.00	0.20	0.14	0.13	0.59	0.83	0.85	0.79	0.60	0.12	0.29
0.00	0.40	0.07	0.06	0.99	0.71	0.82	0.82	0.60	0.31	0.28
0.00	1.00	0.02	0.03	0.46	0.36	0.77	0.74	0.56	0.43	0.30
1.00	0.00	0.00	0.02	0.06	0.09	0.63	0.66	0.56	0.35	0.34

TABLE 7: Species 1 tests, sparse times.

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	0.00	0.441	0.198	0.491	0.495
-1.00	0.50	0.308	0.130	0.170	0.168
-1.00	1.00	0.323	0.167	0.181	0.172
-1.00	2.00	0.351	0.167	0.206	0.200
-0.50	0.20	0.314	0.119	0.175	0.183
-0.50	0.40	0.317	0.154	0.174	0.174
-0.50	0.50	0.321	0.167	0.178	0.175
-0.50	1.00	0.341	0.167	0.201	0.192
0.00	0.00	0.309	0.107	0.208	0.210
0.00	0.05	0.332	0.130	0.187	0.202
0.00	0.10	0.328	0.141	0.179	0.190
0.00	0.20	0.330	0.141	0.186	0.187
0.00	0.40	0.341	0.167	0.203	0.193
0.00	1.00	0.367	0.181	0.218	0.217
1.00	0.00	0.438	0.154	0.277	0.231

Perhaps the best transformation overall is ln(X+.2), partly because it is easier to interpret: there is not much difference among the three or four best. Fig. 1 shows Before and After symmetry plots for this transformation, using all sampling times. This plots $Y_{(Ui)}$-M against M-$Y_{(Li)}$, where $Y_{(Ui)}$ and $Y_{(Li)}$ are, respectively, the i^{th} values above and below the median, and M is a "middle" value, which I took as the point nearest to the 10% trimmed mean in the interval between the (n/2-1) and (n/2+2) (or the (n-1)/2 and (n+3)/2 if n is odd) order statistics. (The median is more commonly used, but its position within the interval formed by the inmost three or four order statistics can affect the appearance of the plot severely.) Except for one Before outlier, the two distributions seem similar in shape and not severely asymmetric. Thus different location estimates should lead to similar tests and estimates of change.

Fig. 1: Symmetry plot for Species 1.



Species 2. Table 8 gives the data, and Table 9 shows preliminary checks on some of the transformations. (The four After times with zeros are omitted when the power is non-positive and the constant is zero.) Although some transformations seem satisfactory, there are standard transformations that do not: X, $X^{.5}$ and ln(X+1) all seem unsuitable.

Results of the two-sample t-test for an effect of the alteration vary from about .06 to .74; other tests are less erratic, but still far from constant (Table 10). However, transformations satisfying the various criteria

TABLE 8: Data for Species 2.

BEFORE

time	I	C	time	I	C
0.641	0.532	0.048	1.495	8.746	6.097
0.644	0.970	0.120	1.514	1.561	1.103
0.721	0.019	0.262	1.533	0.613	1.248
0.723	0.018	0.013	1.552	1.544	0.242
1.189	9.020	33.368	1.571	0.679	1.665
1.208	16.596	27.421	1.590	0.931	0.510
1.227	10.046	95.000	1.609	0.968	0.461
1.257	18.061	8.031	1.628	0.509	0.025
1.284	4.960	12.129	1.648	0.084	0.919
1.309	1.048	1.491	1.667	0.023	0.034
1.342	8.294	70.660	1.686	0.032	0.016
1.361	86.859	21.477	1.724	0.109	0.300
1.380	6.629	19.542	1.735	0.005	0.003
1.399	13.044	4.241	1.743	0.172	0.111
1.418	2.322	2.435	1.762	0.002	0.264
1.437	1.520	2.724	1.781	0.009	0.116
1.459	2.460	0.643	2.515	0.265	0.370
1.478	2.224	0.729	2.879	0.111	0.121

AFTER

time	I	C	time	I	C
4.548	0.948	2.578	6.425	13.661	16.266
4.644	0.030	0.029	6.482	5.112	1.733
4.953	2.167	0.197	6.545	0.236	5.257
5.197	12.083	3.712	6.597	0.004	0.303
5.347	1.597	2.109	6.652	0.048	0.000
5.410	1.351	8.123	6.699	0.007	0.047
5.464	0.087	0.876	7.134	5.573	1.687
5.642	0.000	0.004	7.175	0.164	0.154
5.757	0.000	0.023	7.249	2.453	11.486
5.970	2.176	0.879	7.329	5.654	23.911
6.058	15.739	5.630	7.553	0.081	1.260
6.307	12.633	14.313	7.671	0.000	0.002
6.364	2.129	2.291	7.734	0.015	0.903

are more consistent. The (power, constant) pairs (-1, 1), (-.75, .5 or 1), (-.5, .2 or .5), (-.25, .05 to .2) and (0, 0 to .05) do so here; their t-test p-values vary from .07 to .15, but Wilcoxon ranges from .12 to .17 and biweight from .14 to .2 (except for (0,0), which omits four After observations). The trim-t results are consistently higher. This may be due to the inner half of the After data being more "stretched out" than would be expected from a Normal distribution. All tests seem similarly affected by the After distribution's skewness. Fig. 2 is a symmetry plot for ln(X+.01).

Fig. 2: Symmetry plot for Species 2.

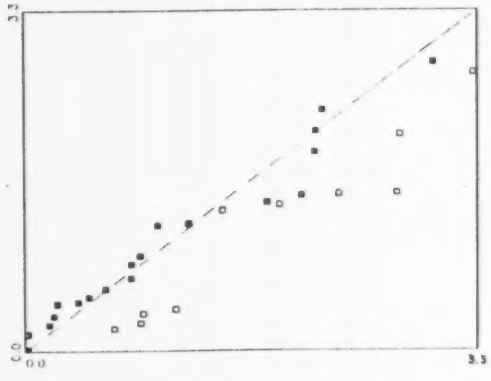


TABLE 9: Preliminary checks for Species 2, all times.

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	0.50	0.81	0.85	-0.08	-0.11	0.63	0.71	1.32	0.30	0.19
-1.00	1.00	-0.84	0.97	-0.34	-0.81	0.71	0.84	1.04	0.31	0.27
-1.00	3.00	-0.24	-0.38	0.01	0.00	0.84	0.85	0.96	0.39	0.70
-0.75	0.20	0.62	0.94	-0.03	-0.03	0.54	0.65	1.08	0.34	0.20
-0.75	0.50	-0.99	0.97	-0.19	-0.54	0.66	0.80	1.00	0.38	0.23
-0.75	1.00	-0.60	-0.71	-0.95	0.32	0.74	0.91	0.96	0.32	0.37
-0.75	3.00	-0.13	-0.34	0.00	0.00	0.90	0.87	0.68	0.47	0.89
-0.50	0.10	0.61	-0.90	-0.03	-0.06	0.51	0.70	0.80	0.31	0.24
-0.50	0.20	0.88	-0.98	-0.11	-0.37	0.60	0.74	0.84	0.37	0.24
-0.50	0.50	-0.66	-0.79	-0.83	0.53	0.71	0.87	0.89	0.39	0.34
-0.50	1.00	-0.32	-0.36	0.09	0.01	0.81	0.88	0.96	0.36	0.56
-0.25	0.05	0.79	-0.92	-0.09	-0.43	0.54	0.68	0.56	0.43	0.31
-0.25	0.10	-0.95	-0.92	-0.31	-0.80	0.62	0.68	0.68	0.43	0.32
-0.25	0.20	-0.64	-0.82	0.97	0.51	0.71	0.79	0.77	0.43	0.37
-0.25	0.50	-0.28	-0.35	0.03	0.01	0.83	0.89	0.96	0.39	0.59
0.00	0.00	0.93	-0.77	-0.39	-0.88	0.55	0.79	0.41	0.45	0.46
0.00	0.01	-0.78	-0.76	0.96	0.53	0.68	0.78	0.44	0.40	0.49
0.00	0.02	-0.64	-0.71	0.59	0.34	0.73	0.78	0.50	0.38	0.51
0.00	0.05	-0.44	-0.66	0.14	0.08	0.80	0.81	0.56	0.41	0.56
0.00	1.00	-0.06	-0.43	0.00	0.00	0.97	0.92	0.64	0.43	0.85
0.50	0.00	-0.05	-0.47	0.00	0.00	0.98	0.92	0.49	0.41	0.68
1.00	0.00	-0.06	-0.40	0.00	0.00	0.97	0.84	1.28	0.32	0.58

Overall abundances vary greatly with time, suggesting a look at "abundant" and "sparse" periods. In this case, these seem to correspond quite well to "seasons": a "Winter" from 0 to .43 seems to contain most of the large values and few of the small. This gives 11 observations in Winter for both Before and After, and 25 Before and 15 After in Summer.

Table 11 gives preliminary checks for Winter. Most transformations with negative powers are acceptable by most of our standards. However, most lead to markedly asymmetric distributions. The positive powers give transformations for which the absolute deviations seem to increase with overall abundance. If this means that effects of the alteration when abundance is high are more influential than those when it is low, this may not be undesirable from some points of view. As it happens, the choice here is not important: all transformations give essentially the same result (Table 12), no significant change.

TABLE 10: Species 2 tests: all times.

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	0.50	0.060	0.154	0.258	0.171
-1.00	1.00	0.076	0.134	0.275	0.141
-1.00	3.00	0.181	0.194	0.262	0.262
-0.75	0.20	0.065	0.125	0.210	0.192
-0.75	0.50	0.069	0.171	0.303	0.140
-0.75	1.00	0.098	0.119	0.297	0.170
-0.75	3.00	0.236	0.292	0.263	0.282
-0.50	0.10	0.075	0.128	0.201	0.158
-0.50	0.20	0.072	0.175	0.291	0.150
-0.50	0.50	0.092	0.134	0.288	0.163
-0.50	1.00	0.142	0.157	0.280	0.211
-0.25	0.05	0.089	0.154	0.222	0.151
-0.25	0.10	0.089	0.171	0.291	0.163
-0.25	0.20	0.102	0.157	0.308	0.173
-0.25	0.50	0.151	0.168	0.290	0.204
0.00	0.00	0.099	0.132	0.174	0.106
0.00	0.01	0.130	0.154	0.208	0.165
0.00	0.02	0.135	0.150	0.254	0.183
0.00	0.05	0.147	0.168	0.284	0.195
0.00	1.00	0.366	0.283	0.257	0.219
0.50	0.00	0.565	0.259	0.240	0.194
1.00	0.00	0.736	0.273	0.203	0.185

TABLE 11: Preliminary checks for Species 2, Winter.

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	1.00	0.70	0.83	0.77	-0.87	0.84	0.66	0.71	0.01	0.34
-1.00	3.00	-0.77	-0.59	0.24	0.45	0.95	0.96	0.47	0.36	0.29
-1.00	10.00	-0.52	-0.35	0.20	0.29	0.97	0.92	0.71	0.43	0.25
-0.75	1.00	-0.99	-0.59	0.43	0.61	0.91	0.96	0.71	0.03	0.30
-0.75	3.00	-0.66	-0.28	0.16	0.45	0.96	0.91	0.47	0.43	0.27
-0.75	10.00	-0.51	-0.40	0.14	0.14	0.97	0.95	0.47	0.36	0.24
-0.50	0.50	-0.88	-0.40	0.27	0.43	0.93	0.91	0.47	0.18	0.27
-0.50	3.00	-0.58	-0.35	0.14	0.24	0.97	0.92	0.47	0.50	0.25
-0.50	10.00	-0.50	-0.38	0.09	0.06	0.96	0.94	0.47	0.50	0.24
-0.25	0.00	-0.79	-0.48	0.14	0.28	0.94	0.92	0.47	0.23	0.25
-0.25	0.01	-0.79	-0.48	0.13	0.28	0.94	0.92	0.47	0.23	0.25
-0.25	0.05	-0.77	-0.48	0.12	0.28	0.95	0.92	0.47	0.23	0.25
-0.25	0.20	-0.73	-0.48	0.10	0.22	0.95	0.92	0.47	0.36	0.25
0.00	0.00	-0.61	-0.38	0.06	0.05	0.96	0.94	0.47	0.50	0.24
0.00	0.01	-0.61	-0.38	0.06	0.05	0.96	0.94	0.47	0.50	0.24
0.00	0.05	-0.61	-0.38	0.06	0.05	0.96	0.94	0.47	0.50	0.24
0.50	0.00	-0.52	-0.38	0.01	0.09	0.95	0.94	0.47	0.36	0.24
1.00	0.00	-0.52	-0.32	0.00	0.05	0.91	0.90	0.47	0.50	0.26

TABLE 12: Species 2 tests, Winter.

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	1.00	0.534	0.693	0.671	0.645
-1.00	3.00	0.605	0.669	0.678	0.664
-1.00	10.00	0.727	0.738	0.667	0.722
-0.75	1.00	0.583	0.669	0.676	0.654
-0.75	3.00	0.652	0.738	0.672	0.684
-0.75	10.00	0.749	0.669	0.669	0.734
-0.50	0.50	0.621	0.716	0.672	0.673
-0.50	3.00	0.697	0.780	0.667	0.703
-0.50	10.00	0.767	0.693	0.675	0.745
-0.25	0.00	0.672	0.716	0.680	0.701
-0.25	0.01	0.672	0.716	0.679	0.701
-0.25	0.05	0.671	0.738	0.676	0.698
-0.25	0.20	0.673	0.738	0.670	0.696
0.00	0.00	0.725	0.760	0.668	0.720
0.00	0.01	0.725	0.760	0.668	0.720
0.00	0.05	0.726	0.760	0.667	0.719
0.50	0.00	0.779	0.780	0.671	0.760
1.00	0.00	0.773	0.799	0.658	0.771

Table 13 gives the preliminary checks for Summer, and Table 14 the results. The acceptable transformations overlap considerably with those for all times together. Also, they give consistent results, except for the trimmed t: a Before-After change that is significant at a level around .05 - .08. The cause of the variable trimmed t results appears to be that the size of the added constant determines whether the inner half of the After data will be relatively near the median (as for a long-tailed distribution), giving low p-values, or stretched out as much as the outer half (like a uniform distribution, although somewhat skewed).

The transformation $-(X+.2)^{-.25}$ seems best overall (I am biased against having deviation size decrease with averages) though we may prefer ln(X+.01) as more interpretable, despite the increase in variation with the averages. Both seem to show no particular pattern in the plot of differences against season (which roughly corresponds to decreasing overall abundance), as shown for ln(X+.01) in Fig. 3. They give similar results: conventional non-significance overall, though near enough to borderline to warrant a closer look; clear non-significance in Winter, with even a hint of an increase at the Impact area, although the Before data are limited to a 13-week period; and a significant decrease (or one near enough to be taken seriously, by most standards) in Summer.

TABLE 14: Species 2 tests, Summer.

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	0.10	0.073	0.055	0.083	0.067
-1.00	0.50	0.057	0.069	0.176	0.079
-0.75	0.10	0.064	0.058	0.100	0.082
-0.75	0.50	0.062	0.065	0.185	0.085
-0.50	0.05	0.068	0.055	0.081	0.075
-0.50	0.20	0.058	0.073	0.165	0.083
-0.25	0.00	0.054	0.039	0.086	0.031
-0.25	0.01	0.077	0.052	0.038	0.043
-0.25	0.05	0.062	0.062	0.118	0.085
-0.25	0.20	0.065	0.069	0.183	0.091
0.00	0.00	0.047	0.049	0.082	0.043
0.00	0.01	0.065	0.052	0.115	0.075
0.00	0.02	0.065	0.049	0.131	0.085
0.00	0.05	0.067	0.069	0.160	0.093
0.00	0.10	0.071	0.073	0.179	0.098
0.50	0.00	0.102	0.065	0.150	0.133
1.00	0.00	0.192	0.115	0.211	0.321

Fig. 3: Species 2 Differences vs. Season

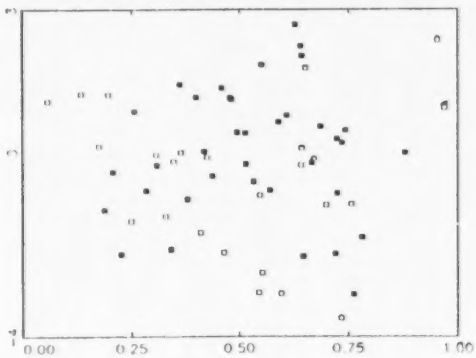


TABLE 13: Preliminary checks for Species 2, Summer.

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	0.10	0.38	0.63	-0.22	-0.11	0.41	0.66	0.69	0.31	0.35
-1.00	0.50	0.41	0.28	0.72	0.30	0.63	0.64	0.93	0.36	0.20
-0.75	0.10	0.38	0.40	-0.34	-0.60	0.45	0.52	0.50	0.45	0.30
-0.75	0.50	0.41	0.22	0.48	0.12	0.66	0.67	0.93	0.29	0.20
-0.50	0.05	0.35	0.47	-0.28	-0.41	0.42	0.54	0.47	0.39	0.32
-0.50	0.20	0.39	0.24	0.97	0.50	0.59	0.53	0.72	0.41	0.22
-0.25	0.00	0.19	0.86	-0.02	-0.01	0.22	0.69	0.44	0.26	0.44
-0.25	0.01	0.28	0.60	-0.17	-0.21	0.34	0.61	0.90	0.12	0.36
-0.25	0.05	0.35	0.31	-0.52	-0.93	0.50	0.48	0.50	0.27	0.27
-0.25	0.20	0.38	0.23	0.55	0.18	0.63	0.62	0.80	0.34	0.22
0.00	0.00	0.25	0.52	-0.24	-0.43	0.37	0.50	0.44	0.11	0.32
0.00	0.01	0.30	0.32	-0.52	-0.89	0.47	0.46	0.38	0.18	0.29
0.00	0.02	0.31	0.30	-0.73	0.91	0.51	0.46	0.50	0.31	0.27
0.00	0.05	0.34	0.18	0.83	0.28	0.57	0.45	0.50	0.43	0.24
0.00	0.10	0.35	0.19	0.47	0.13	0.62	0.54	0.66	0.32	0.23
0.50	0.00	0.17	0.19	0.02	0.00	0.58	0.66	0.66	0.16	0.28
1.00	0.00	0.00	0.18	0.00	0.00	0.30	0.54	0.88	0.08	0.49

TABLE 15: Data for Species 3.

BEFORE

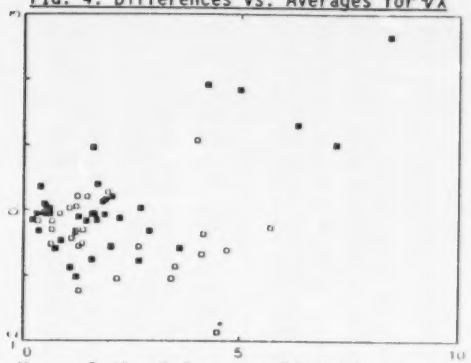
time	I	C	time	I	C
0.562	7.375	7.234	1.495	0.509	3.011
0.641	2.388	2.609	1.514	0.339	0.419
0.644	2.549	2.726	1.533	0.237	0.308
0.721	4.353	1.275	1.552	0.319	0.042
0.723	3.600	3.149	1.571	0.035	0.259
1.189	48.513	32.010	1.590	4.610	3.792
1.208	95.000	50.123	1.609	7.529	9.404
1.227	59.158	44.649	1.628	3.352	3.642
1.257	11.007	15.227	1.648	0.193	1.057
1.284	27.064	10.728	1.667	0.396	1.203
1.309	34.939	16.510	1.686	1.493	1.760
1.342	4.654	5.212	1.724	0.417	2.320
1.361	3.039	5.315	1.735	0.192	0.234
1.380	2.632	3.163	1.743	0.383	0.356
1.399	1.866	2.358	1.762	0.073	0.110
1.418	1.465	3.884	1.781	0.015	0.074
1.437	5.217	9.350	2.515	3.913	3.292
1.459	2.579	2.814	2.707	0.274	0.185
1.478	3.661	2.300	2.879	0.299	0.235

AFTER

time	I	C	time	I	C
4.548	2.476	1.887	6.425	0.451	3.649
4.644	0.795	1.784	6.482	0.247	0.646
4.953	9.510	15.629	6.545	15.770	18.839
5.197	31.010	34.075	6.597	0.999	2.428
5.347	1.546	1.432	6.652	12.619	29.479
5.410	1.053	1.893	6.699	2.693	7.268
5.464	0.135	0.798	7.134	14.308	19.925
5.642	1.500	2.364	7.175	19.374	25.146
5.699	5.673	8.689	7.249	20.595	12.051
5.757	0.058	0.170	7.329	8.379	15.565
5.970	1.823	1.313	7.553	1.147	1.107
6.058	0.651	0.752	7.671	4.354	3.298
6.307	1.204	2.617	7.734	0.320	0.550
6.364	1.107	1.877			

Species 3. The data are in Table 15 and examples from the first run of the preliminary checks are in Table 16. These examples, and more detailed tables, strongly suggest a problem with serial correlation. (The rank version of von Neumann's test seems to have greater power, as Bartels (1982) claimed it would for non-Normal distributions.) In addition, no transformation seems additive. This appears to be due to the six high density sampling times in the Before period, when Impact densities were, with one mild exception, greater than Control densities, while this ordering is reversed for most other sampling times. Fig. 4 shows the plot of the differences against the averages when the square root transformation is used.

FIG. 4: Differences vs. Averages for $\sqrt{X}$



Many of the Before sampling times are quite close together (most are a week apart), so serial correlation is not surprising and some adjustment seems needed. I have not attempted any serious modelling here. I have assumed that the centered differences, $d_{ij}=D_{ij}-E(D_{ij})$ $(D_{ij}=Y_{ij1}-Y_{ij2})$, obey $d_{ij}=r_{ij}d_{i(j-1)}+\varepsilon_{ij}$ where the ε_{ij}'s are independent errors and r_{ij} decreases linearly with the timegap, $t_{ij}-t_{i(j-1)}$. I chose the estimated line by an ad hoc argument: I assumed that $r_{ij}=0$ if the timegap is greater than .15 (this is about eight weeks, and contains the period of high density Before observations), and that the sample autocorrelation estimated r_{ij} for a timegap of about .019, the median (and mode) of the Before timegaps. This gives $r_{ij}=1.15r(1-T/.15)$, where T is the timegap. If $r_{ij}>0$, I replaced D_{ij} by $(D_{ij}-r_{ij}D_{i(j-1)})$, divided by $(1-r_{ij})$ so the mean would be unchanged.

Table 17 shows the preliminary checks for these adjusted samples. The serial correlation problem seems now to be minor, but additivity remains elusive. A plot like Fig. 4, but using the adjusted data, looks very like Fig. 4. (Trends in the absolute deviations, portrayed in "vslope" and "vspear", may be due to trends in the mean.) It is possible that, without the alteration, Impact is more popular than Control in some conditions and less popular in others. I therefore broke each data set into two, defining "conditions" by time-of-year (Summer between .33 and .9) rather than by "abundance" (e.g. Control>10), though the answers are similar.

Tables 18 and 19 give preliminary checks on the adjusted data for Summer and Winter respectively. Even though there are only six

TABLE 16: Preliminary checks for Species 3, all times.

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	0.00	0.00	0.02	-0.00	-0.00	0.74	0.17	1.57	0.03	0.59
-1.00	1.00	0.33	0.09	-0.13	-0.10	0.36	0.07	0.86	0.44	0.53
-1.00	3.00	0.25	0.25	-0.97	0.42	0.28	0.07	0.79	0.37	0.43
-1.00	5.00	0.16	0.28	0.43	0.04	0.22	0.07	0.79	0.46	0.40
-0.50	0.00	0.00	0.03	-0.00	-0.00	0.90	0.15	0.70	0.08	0.61
-0.50	1.00	0.19	0.13	-0.73	0.63	0.27	0.10	0.79	0.45	0.44
-0.50	3.00	0.07	0.21	0.17	0.03	0.17	0.05	0.72	0.44	0.36
-0.50	5.00	0.02	0.17	0.02	0.00	0.11	0.04	0.65	0.50	0.34
0.00	0.00	0.03	0.02	-0.05	-0.27	0.63	0.23	0.88	0.41	0.43
0.00	0.10	0.05	0.04	-0.95	-0.97	0.28	0.18	0.72	0.48	0.41
0.00	0.50	0.02	0.09	0.11	0.15	0.12	0.07	0.72	0.37	0.34
0.00	1.00	0.01	0.14	0.01	0.02	0.09	0.04	0.65	0.28	0.31
0.50	0.00	0.00	0.09	0.00	0.00	0.01	0.07	0.77	0.22	0.26
1.00	0.00	0.00	0.24	0.00	0.00	0.00	0.04	2.04	0.49	0.29

TABLE 17: Preliminary checks for Species 3, adjusted for correlation, all times

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	3.00	0.30	0.30	-0.86	0.39	0.53	0.27	0.09	0.51	0.53
-0.50	1.00	0.23	0.19	-0.61	0.87	0.54	0.39	0.09	0.49	0.53
-0.50	3.00	0.12	0.39	0.29	0.08	0.54	0.49	0.15	0.45	0.49
0.00	0.00	0.03	0.02	-0.05	-0.27	0.63	0.23	-0.06	0.41	0.43
0.00	0.10	0.06	0.09	-0.81	0.99	0.54	0.40	0.09	0.37	0.49
0.00	0.50	0.04	0.17	0.29	0.22	0.57	0.69	0.18	0.42	0.48
0.00	1.00	0.03	0.22	0.07	0.06	0.56	0.68	0.22	0.41	0.47
0.50	0.00	0.00	0.14	0.00	0.07	0.50	0.91	0.38	0.40	0.48
0.50	0.50	0.00	0.26	0.00	0.02	0.46	0.89	0.39	0.34	0.49
1.00	0.00	0.00	0.28	0.00	0.01	0.44	0.89	0.45	0.47	0.50

TABLE 18: Preliminary checks for Species 3, adjusted for correlation, Summer

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	3.00	0.91	-0.67	0.77	0.69	0.51	0.20	0.04	0.23	0.72
-0.50	1.00	0.82	-0.98	-0.88	-0.93	0.50	0.28	0.03	0.15	0.75
-0.50	3.00	-0.86	-0.46	0.44	0.29	0.51	0.23	0.06	0.13	0.75
0.00	0.00	0.31	0.25	-0.01	-0.09	0.83	0.47	-0.18	0.36	0.71
0.00	0.10	0.71	0.44	-0.36	-0.60	0.56	0.35	-0.04	0.34	0.72
0.00	0.50	0.99	-0.78	0.82	0.86	0.51	0.31	0.04	0.16	0.77
0.00	1.00	-0.87	-0.59	0.51	0.35	0.51	0.25	0.06	0.18	0.76
0.50	0.00	-0.64	-0.80	0.32	0.33	0.51	0.35	0.05	0.12	0.79
0.50	0.50	-0.45	-0.36	0.12	0.18	0.51	0.38	0.09	0.16	0.81
1.00	0.00	-0.04	-0.39	0.00	0.01	0.48	0.47	0.13	0.05	0.90

TABLE 19: Preliminary checks for Species 3, adjusted for correlation, Winter

power	const	Tukey	Spear	vslope	vspear	von N	rank vN	shape	Gupta	trend
-1.00	3.00	0.97	-0.90	-0.01	-0.10	0.56	0.32	-0.11	0.50	0.37
-0.50	1.00	0.87	1.00	-0.00	-0.05	0.57	0.41	-0.12	0.50	0.43
-0.50	3.00	0.84	1.00	-0.00	-0.05	0.56	0.41	-0.11	0.50	0.45
0.00	0.00	0.62	1.00	-0.05	-0.09	0.59	0.41	-0.14	0.50	0.62
0.00	0.10	0.62	1.00	-0.05	-0.09	0.59	0.41	-0.14	0.50	0.62
0.00	0.50	0.61	1.00	-0.06	-0.09	0.59	0.41	-0.14	0.50	0.63
0.00	1.00	0.60	1.00	-0.07	-0.09	0.59	0.41	-0.13	0.50	0.64
0.50	0.00	0.23	0.37	-0.64	-0.95	0.59	0.73	-0.11	0.50	0.95
0.50	0.50	0.23	0.37	-0.65	-0.95	0.59	0.73	-0.11	0.50	0.94
1.00	0.00	0.04	0.20	0.47	0.70	0.52	0.32	-0.02	0.50	0.54

Before observations in Winter, many transformations are ruled out by trends in the sizes of the deviations. There are 32 Before observations in Summer, but there is a much broader range of acceptable transformations: Impact and Control differ little in this period, and it may also be that, for very small observations, transformations involving a relatively large added constant behave rather alike.

If there is a "season-x-site" interaction, it may not be the case that the same transformation is appropriate for both seasons. However, the square root transformation seems to be satisfactory in both seasons here. More reassuringly, the tests for an effect of the alteration give very similar results for acceptable (and many unacceptable) transformations for all times (Table 20), Summer alone (Table 21) and Winter alone (Table 22).

TABLE 20: Species 3 tests (cf Table 17).

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	3.00	0.040	0.011	0.015	0.008
-0.50	1.00	0.036	0.010	0.010	0.004
-0.50	3.00	0.018	0.008	0.007	0.011
0.00	0.00	0.041	0.016	0.012	0.033
0.00	0.10	0.017	0.007	0.004	0.010
0.00	0.50	0.014	0.005	0.007	0.007
0.00	1.00	0.012	0.006	0.006	0.010
0.50	0.00	0.016	0.017	0.030	0.050
0.50	0.50	0.018	0.020	0.027	0.057
1.00	0.00	0.048	0.033	0.089	0.077

TABLE 21: Species 3 tests (cf Table 18).

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	3.00	0.036	0.014	0.017	0.009
-0.50	1.00	0.032	0.011	0.021	0.006
-0.50	3.00	0.028	0.016	0.025	0.011
0.00	0.00	0.057	0.043	0.021	0.060
0.00	0.10	0.033	0.029	0.018	0.029
0.00	0.50	0.020	0.009	0.014	0.010
0.00	1.00	0.021	0.011	0.017	0.011
0.50	0.00	0.019	0.011	0.008	0.034
0.50	0.50	0.025	0.017	0.023	0.019
1.00	0.00	0.072	0.018	0.113	0.027

TABLE 22: Species 3 tests (cf Table 19).

power	const	t-test	Wilc	trim-t	biwt-t
-1.00	3.00	0.035	0.034	0.119	0.053
-0.50	1.00	0.022	0.034	0.083	0.031
-0.50	3.00	0.016	0.019	0.063	0.020
0.00	0.00	0.008	0.010	0.037	0.008
0.00	0.10	0.008	0.010	0.036	0.008
0.00	0.50	0.008	0.007	0.035	0.008
0.00	1.00	0.008	0.007	0.033	0.007
0.50	0.00	0.006	0.005	0.033	0.005
0.50	0.50	0.006	0.005	0.034	0.005
1.00	0.00	0.013	0.004	0.073	0.009

Thus this species seems to have been reduced by the alteration. However, the evidence is stronger for Summer, when the two sites seem to have been quite similar (Before) and there are many observations, than for Winter, when the relation between the sites seems to have been different but there are only six Before observations, all within a seven week period.

4. Discussion.

There are three obvious directions for further work: better testing, better things than testing, and lessons to be applied in the design of assessment programs.

With respect to the first, there is a laundry list of desirable extensions. A proper plan for dealing with correlated data is high on this list. However it seems unlikely to be accomplished without a model involving more than half a dozen parameters, and requiring larger and longer (in time) datasets than are often customary.

Multivariate procedures are another item. It will have been noticed that Species 2 and 3 were both dense at about the same times. If they are similar biologically (or suspected to be so), we may get more reliable results, and perhaps some idea about the mechanisms of any effect, by analyzing them together. Unfortunately, knowledge of many populations is such that it is easy to "suspect" links between large numbers of species, and difficult to rate the credibility of these suspicions. Data-based methods for linking different species would be helpful.

Mention of many species raises the problem of multiple testing. In many assessment problems there may be dozens or even hundreds of species to consider: all are sampled the same way, so the "extra" data are essentially free. Even if no species is affected, the probability that at least one will be "significant" may be much greater than the nominal level. The standard answer, which amounts largely to using very small test levels, is unsatisfactory: Normal and t-approximations can't be trusted at these levels, and the variances and sample sizes are such that even severely affected species are unlikely to register at them.

Can we do better than significance testing? That many resource managers and biologists seem to want, and be content with, a test of significance appears sometimes to be a remarkable public relations achievement. The meaning of a p-value is not easy to grasp, even in simple cases where the model is known. It may not be understood by many of those who ask for it. There are plenty of logical problems, as the Bayesians have pointed out (e.g. Edwards, Lindman and Savage 1963, Lindley 1972).

Arguments can be made for impact assessment testing. An obvious one is that environmental protection laws often imply tests rather than estimates: alterations are to make no change in the "balanced indigenous" community. Another is that estimation may require larger sample sizes. It may be more economical to test first and then decide whether further lab and field studies are justified. Bayesians have pointed out that the p-values from one-sided tests may be interpretable Bayesianly, if the prior is diffuse. (It surely should be in much of this work, and when environmentalists and industrialists each have sharp, but sharply different, priors, the p-value seems useful as an objective compromise.) Also, difficult as testing is, estimation may be much harder, both to do and to explain.

But this last objection comes near to violating Tukey's principle, "better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can often be made precise." The right question is surely "Do the effects matter?" This is vague because it involves a value judgment, but statisticians can help individuals to decide for themselves by answering the nearby question "How big are the effects?"

This question is easy to answer, at least approximately, from the testing work only if Control is believed to be unchanged and the transformation used in the tests implies the units we want. A change in X gives us the absolute change; in ln(X) the percentage change; in X^{-1} the change in space needed per animal. Added constants might be justified on the grounds that $E(X+c)^b \approx (EX)^b$ approximately, using Jensen's inequality. But powers other than -1, 0 and 1 seem hard to interpret, although they may well be justified in terms of "damping". Also, if the units we want do not correspond to the transformation we used, there is no simple back-conversion to help: the estimated size of the change will involve the actual sample abundances, which will usually depend on the choice of sampling times.

An alternative is to attack estimation directly. We are immediately faced with the problem of choosing units, the possibility that the effect may vary (e.g. with time-of-year), and the difficulty of presenting estimates (with confidence intervals, etc.) in a form quickly grasped by non-scientists. But these questions suggest a useful approach. This is to use the Control (or several Controls) as predictors, and to present the assessment in the form of a plot showing, for each Control value, the Impact value expected in the Before period and the Impact value expected in the After period. (One can also plot the observed After value at Impact against the Before value that would be predicted from the observed After values at Control.) Thus the Control stands for environmental conditions which might determine the direction and size of the change at Impact. This approach was successfully used by Mathur, Robbins and Purdy (1980), but I have not seen it elsewhere.

This approach requires the calculation of regressions, of Impact against Control(s), for Before and After. The problems arising in testing arise here too, often more severely. Corresponding to transformation choice is the choice of a model. This latter choice is wider, though: there are natural models that allow Control to be more abundant than Impact in some conditions and less abundant in others. Also there are new possibilities, like isotonic regression (Barlow et al 1972) or projection pursuit regression (Friedman and Stuetzle 1981), that allow us to evade a particular choice. These bring more problems, some of them severe, such as estimating variance (especially if it varies) and detecting and allowing for serial correlation.

Finally, two design considerations should be mentioned. First, it may be inadequate to decide on a sample size only by the usual calculation, $n = 2s^2(z_\alpha - z_\beta)^2/d^2$, where α is the test level, d is a difference it is desired to detect with probability β, s^2 is a variance estimate, and z_α and z_β are the upper a and b points of the N(0,1) distribution. (The units of d and s^2 are those of a transformation applied to some

preliminary data used to obtain s^2.) While such calculations need to be done, we need to anticipate uncertainty about transformations, the need to split the sample into two or more parts, serial correlation, and the possibility of year-to-year variation.

Secondly, the estimation approach makes explicit what was only implicit in testing. This is that the "control" is not a control in the sense used in experiments. In the testing problem, we think of it as "tracking" Impact, if only we could find the right transformation. But the estimation approach described above makes it clear that it is a predictor. Thus it is not essential that Control be "similar" to Impact, only that it predict it well Before the alteration. Thus several Controls, serving complementary purposes, might be useful, as might be observations at "Control" on species other than the target one.

Acknowledgement: I thank the Marine Review Committee for permission to use the data in Section 3. MRC does not necessarily agree with any of the opinions in this article.

REFERENCES

Anderson, T. W. 1970. The Statistical Analysis of Time Series. Wiley, New York.

Antille, A. G., G. Kersting and W. Zucchini. 1982. Testing Symmetry. Journal of the American Statistical Association, 77; 639-646.

Atkinson, A. C. 1985. Plots, Transformations and Regression. Oxford University Press, New York.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner and H. D. Brunk. 1972. Statistical Inference Under Order Restrictions. Wiley, New York.

Bartels, R. 1982. The Rank Version of von Neumann's Ratio Test for Randomness. Journal of the American Statistical Association, 77; 40-46.

Berry, D. A. Logarithmic Transformations in ANOVA. Biometrics, 43; 439-456.

Box, G. E. P., and D. R. Cox. 1964. An Analysis of Transformations. Journal of the Royal Statistical Society, Series 'B', 26; 211-252.

Box, G. E. P., and W. J. Hill. 1974. Correcting Inhomogeneity of Variances with Power Transformation Weighting. Technometrics 16; 385-389.

Box, G. E. P., and G. C. Tiao. 1965. A Change in the Level of a Non-Stationary Time Series. Biometrika, 52; 181-192.

Box, G. E. P., and G. C. Tiao. 1975. Intervention Analysis with Applications to Economic and Environmental Problems. Journal of the American Statistical Association, 70; 70-79.

Edwards, W., H. Lindman and L. J. Savage. 1963. Bayesian Statistical Inference for Psychological Research. Psychological Review, 70; 193-242.

Friedman, J. and W. Stuetzle. 1981. Projection Pursuit Regression. Journal of the American Statistical Association, 76; 817-823.

Gupta, M. K. 1967. An Asymptotically Non-Parametric Test of Symmetry. Annals of Mathematical Statistics, 38; 849-866.

Kafadar, K. 1982. Using Biweight Estimates in the Two-Sample Problem. Part 1: Symmetric Populations. Communications in Statistics - Theory and Methods, 11; 1883-1901.

Kafadar, K. 1983. The Efficiency of the Biweight as a Robust Estimator of Location. Journal of Research of the National Bureau of Standards, 88; 105-116.

Lehmann, E. L. 1975. Non-Parametrics: Statistical Methods Based on Ranks. Holden Day.

Lindley, D. V. 1972. Bayesian Statistics: A Review. S. I. A. M., Philadelphia.

Mathur, D., T. W. Robbins and E. J. Purdy. 1980. Assessment of Thermal Discharges on Zooplankton in Conowingo Pond, Pennsylvania, Canadian Journal of Fisheries and Aquatic Science 37; 937-944.

Posten, H. O. 1978. The Robustness of the Two-Sample t-test Over the Pearson System. Journal of Statistical Computation and Simulation, 6; 295-311.

Posten, H. O. 1979. The Robustness of the One-Sample t-test Over the Pearson System. Journal of Statistical Computation and Simulation, 9; 133-149.

Posten, H. O. 1982. Two-sample Wilcoxon Power Over the Pearson System and Comparison with the t-test. Journal of Statistical Computation and Simulation, 16; 1-18.

Reinsel, G. C., and G. C. Tiao. 1987. Impact of Chlorofluoromethanes on Stratospheric Ozone. Journal of the American Statistical Association, 82; 20-30.

Stewart-Oaten, A., W. W. Murdoch and K. R. Parker. 1986. Environmental Impact Assessment: Pseudo-Replication in Time? Ecology 67(4); 929-940.

Stewart-Oaten, A. 1986. Assessing Local Impacts: Progress and Some Problems. Oceans '86 Conference Record, Vol 3, pp 964-973. Available from: IEEE Service Center, 445 Hoes Lane, Piscataway, N. J., 08854; or from: Marine Technology Society, 2000 Florida Avenue, N. W., Suite 500, Washington, D. C.

Stigler, S. M. 1976. The Effect of Sample Heterogeneity on Linear Functions of Order Statistics, with Applications to Robust Estimation. Journal of the American Statistical Association, 71; 956-960.

Tukey, J. W. 1949. One Degree of Freedom for Non-Additivity. Biometrics, 5; 232-242

von Neumann, J. 1941. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. Annals of Mathematical Statistics, 12; 367-395.

Yuen, K. K. and W. J. Dixon. 1973. Approximate Behavior and Performance of the Two-Sample Trimmed t. Biometrika, 60; 369-374.

COMPARISONS WITH BACKGROUND ENVIRONMENT: STRATEGIES FOR DESIGN

Bruce Peterson
CH2M HILL, 777 108th Avenue NE, Bellevue, WA 98004

INTRODUCTION

Current federal and state environmental regulations have greatly increased the number of environmental studies being performed each year in the United States. These studies range from investigations of Superfund sites to the establishment of monitoring systems for RCRA-permitted facilitates. A common component of all of these studies is the comparison of the sample values of environmental parameters with background samples or regulatory thresholds.

Design of environmental sampling strategies in a potentially litigious regulatory context presents many new challenges for the statistician. Among these are the traditional challenges of adequate problem definition, newer challenges of maintaining data quality when multiple organizations handle the samples, and the challenges in the rapidly developing field of chemometrics (the study of the statistical properties of chemical measurements).

Studies involving statistical comparisons have an initial phase during which the objectives of the study and the contrasts of interest are refined and agreed upon. Regulatorily mandated environmental studies, particularly of Superfund sites, often involve multiple actors in defining the study objectives. Each of these actors--the regulatory agencies, the potentially responsible parties, and the public--have potentially disparate agendas. Each agenda can be translated into statistical design goals that may be incompatible with those derived from other agendas. Frequently the actors are statistically unsophisticated: the definition of study objectives therefore becomes an involved process of both education and political maneuvering.

Even with an agreed-upon set of study objectives, the regulatory context of a study results in a high probability of litigation. This places an obligation on the study team to carefully document and track sampling activities and the movement of samples through the analytical process. Well-constructed data bases are a necessity.

The data base should include provisions for recording progress of samples through the analysis, field changes to the sample plan, variations in the analytical protocol, pointers to the original documents, and the values of the environmental parameters.

The data base will ultimately provide the information necessary for statistical analysis and comparison of the environmental parameters. The challenges of the analysis will include the mundane: obtaining a "final" data set, and the complex: performing comparisons with multivariate and incomplete data. In addition the data base must maintain information on the quality of the chemical analysis for incorporation into the estimate of the overall quality of analysis.

The chemometric measures of quality stored in a data base enable multiple uses of the data. These measures are available from the analytical laboratories, allowing data of differing quality to be used in analyses where appropriate. As an analytical lab is capable of generating a large number of quality measures, these are best transferred electronically.

STRATEGIES FOR DESIGN: THE LOVE CANAL PILOT STUDY

Many of these challenges have been encountered with the Love Canal Superfund site. Construction on the Love Canal was started in the 1890s but never completed. The canal was used as a chemical waste dump from 1942 to 1953 and was then filled and covered. The Love Canal area developed as an urban neighborhood until significant contamination of the environment was recognized in 1978. Extensive remediation to isolate and contain the wastes in the canal was performed in 1979 and additional remediation to clean the surrounding area is continuing. The Habitability Study, due for completion in May 1988, is designed to provide information on the habitability of the emergency declaration area (EDA) surrounding the canal.

The technical goals of the study were developed by a Technical Review Committee (TRC) established by the U.S. EPA to oversee and coordinate the habitability study. TRC members represent the U.S. EPA, the N.Y. State Department of Health, the U.S. Department of Health and Human Services/Centers for Disease Control, and the N.Y. State Department of Environmental Conservation. The TRC in turn formed a scientific advisory panel to formulate and recommend habitability criteria.

The scientific panel recommended that a habitability study be formulated as a statistical comparison between each EDA neighborhood and similar areas not influenced by a hazardous waste site in the Niagara Falls/Buffalo area. The comparison was to be based on the chemical concentrations of 8 indicator chemicals (LCICs), chosen for mobility and specificity to the canal. The comparison was to provide a 90 percent probability of detecting an "order of magnitude" difference in median values with 95 percent confidence against the null hypothesis that no difference in median values

existed. The sample design was developed after internal discussion and a peer review. The full criteria can be found in Love Canal Emergency Declaration Area: Proposed Habitability Criteria, NYSDOH and DHHS/CDC 1986.

This degree of specificity is unusual in Superfund-driven environmental studies. However, the extensive discussions that preceded agreement on these criteria served to develop a common expectation as to what the habitability study should provide.

The habitability study, like most environmental studies, required estimates of LCIC concentration variability to design the sampling and analysis plan. In addition, since previous studies had found the indicator chemicals to have a high percentage of undetectable concentrations, greater sensitivity of the chemical analysis was needed.

A pilot study was designed and implemented to test the field operations required to collect the samples, test a new analysis method for estimating LCIC concentrations, and identify and quantify sources of variability in the concentration estimates.

The Love Canal pilot study illustrated difficulties common to many large environmental studies, one example being the tracking of samples from field collection through analysis. Although a data base was established to track samples, the number of changes that occurred in the field required considerable effort to trace during later stages of the study. Environmental studies should establish sample tracking and documentation as priority task.

Laboratory variability, not generally quantified in many environmental studies, was found to be significant during the pilot. Laboratory operations were closely tracked as part of the pilot, yielding quantitative measures of variability not commonly available through EPA's contract laboratory program. This close tracking was continued for the full study, as the detailed information provides insight about the quality of the data.

The pilot provided information on the sources of variability and bias in the laboratory estimates of concentration. Interlaboratory variability was found to be of similar magnitude as intralaboratory variability. Biases occurred among laboratories from several causes ranging from contamination levels in different lots or brands of reagents to differences in exact analytical protocols among laboratories.

An analytical method specific for LCICs was developed to achieve sufficient sensitivity to increase the number of samples with detectable concentrations. Criticism of earlier studies about the sensitivity of the analytical methods led to an effort to estimate the "detection limits" for the analytical method.

The analytical method proposed for the habitability study uses a gas chromatograph/mass spectrometer (GC/MS) instrument similar to that used for most sensitive organic analyses of environmental samples. The method differs from the usual broad scan for a range of compounds. In this method the instrument was programmed to search only for LCICs by a method known as single ion monitoring (SIM).

The GC/MS instrument is sensitive, providing quantitative estimates of very-low-level concentrations (parts per billion range), and accurate, providing a high level of confidence as to the identity of the compound. The proposed method increased the sensitivity of the instrument by operating it in the SIM mode. This concentrated the instruments' scan time on the LCICs.

Estimating the "detection limit" for the GC/MS SIM proved difficult as the chemometric literature has not dealt extensively with detection limits for multivariate instruments. A sample specific detection limit cannot be estimated for the GC/MS SIM method. However, a method detection limit may be estimated. The method detection limit is the concentration at which samples with a known concentration of a compound would be differentiated from a blank 95 percent of the time.

An estimator for the method detection limit is difficult to develop as identification and quantification are confounded in the GC/MS instrument. Identification and quantification is a multivariate process where a compound is not quantified unless it is first identified. Generally the ability to identify a compound is lost at concentrations greater than those at which quantification could occur. Identification often fails because of chemical interferences in the chromatogram at low concentrations. These interferences obscure the ion area peaks associated with the target compound.

Standard estimators of detection limits (Long and Winefordner 1983, Currie 1968, Hubaux and Vos 1968) are defined in terms of instruments with univariate responses. Since the GC/MS is a multivariate instrument, these methods are not directly applicable. New estimators are required for detection (or identification) limits for the GC/MS. These are being developed in the full habitability study.

SUMMARY AND CONCLUSION

Environmental studies are increasingly being mandated by regulatory concern. In this potentially litigious context, exceptional efforts are required to produce

sound statistical designs. The objectives of each study must be clearly defined and conflicting objectives identified and mediated before the study can be successful.

A good data base management system is a prerequisite for a successful environmental study. This system should include the personnel, software, and hardware to track not only the parameter estimates but sample locations, sequences, quality assurance parameters, and pointers to original documents. With environmental studies typically involving a number of organizations, inadequacies of data management will cause difficulties in data defensibility and interpretation.

Finally, a good understanding of the analytical process that produces the parameter estimates is required. Experience with the analytical techniques used at Love Canal has shown that basic concepts such as "Detection Limit" are not as clearly defined as may be necessary. Statisticians must expect to work closely with chemists and other experts to evaluate the impact of chemical concentration estimators on study objectives.

REFERENCES

Currie, L. Limits for Qualitative Detection and Qualitative Determination. Analytical Chemistry, Vol. 40, No. 3. 1968. Pp. 586-593.

Hubaux, A., and G. Vos. Decision and Detection Limits for Linear Calibration Curves. Analytical Chemistry, Vol. 42, No. 8. 1968. Pp. 849-855.

Long, G. L., and J. D. Winefordner, Limit of Detection: A Closer Look at the IUPAC Definition. Analytical Chemistry, Vol. 55, No. 7. 1983. Pp. 712-724.

New York State Department of Health and U.S. Department of Health and Human Services/Centers for Disease Control. Love Canal Emergency Declaration Area: Proposed Habitability Criteria. 1986.

STATISTICAL COMPARISONS OF SITES:

DISCUSSION

R. Clifton Bailey
Environmental Protection Agency, WH586, 401 M. Street, SW
Washington, DC 20460

The common the problem in both papers is the choice of a comparison areas or sites. The approaches differ in the use of sites similar in time versus similar in location. These choices are made for practical considerations. In both cases, traditional statistical techniques are stretched to fit the situation.

Both papers emphasize statistical methods adapted from traditional work for designed experiments, and emphasize making the comparison sites similar. There are, of course, always problems with finding similar sites. Whatever the choice, in controversial situations, there is always an advocate to find fault with the choice. The problem is especially troublesome since traditional safeguards such as randomization are not available.

Since it is difficult to find an ideal sit for comparison, we might examine the fundamental question: Do we want just one area for comparison? If more than one area is used for comparison, then how do we make use of the information from multiple sites?

If more than one comparison site is used, then we could establish relationships between factors thought to be important, and use the results for adjustment. Often this ploy is used in observational studies when the conditions do not permit the usual design interventions such as the random assignment of treatment and control sites.

To establish relationships requires a range of values for variables of potential concern. This means that we seek a range of site conditions which can be used to establish the adjustment relationships rather than trying to find identical sites. Since identical sites don't exist, we need enough information to construct a hypothetical site analytically though computed adjustments. Consequently more than one control site must be used unless there is an obvious control or reference that cannot be questioned.

In effect you compute the ideal control site rather than find an almost ideal physical site. The control sites serve to calibrate your model so that you can adjust for other factors.

In these situations, where traditional statistical design don't apply, there will always be factors which will differ between sites and which someone will point to as an explanation for any difference found by traditional statistical tests.

If we include a reasonable set of concomitant information to make our adjustments, the analysis must be more complex.

Traditional statistical designs usually place a great deal of emphasis on making the analysis simple. When it is not possible to use traditional statistical design to make the analysis easy, then we must consider more complex analyses to circumvent design deficiencies.

Design deficiencies may be circumvented by taking a global approach in which we use information from several studies jointly for common elements when the information in individual studies is inadequate.

Success with the global approach requires a modeling effort based on a subject matter understanding far beyond the models of traditional statistical methods.

Similar ideas were expressed W. G. Cochran(1965). The Planning of Observational Studies of Human Populations, Journal of the Royal Statistical Society, Series A, Vol. 128, Pt. 2, pp 234–265. This is a good source of ideas for dealing with studies that are outside the realm of traditional statistical designs. For example, Cochran says:

> Planning of observational studies are found in subject matter departments not in statistics departments.
>
> Effective planning of observational studies calls for considerable mastery of subject-matter field.
>
> In setting up comparisons, the investigator makes do with comparisons that are far from ideal.
>
> The time and resources for analysis of results are often underestimated.
>
> To handle disturbing variables-- list variables for which matching or adjustment is
> 1) required,
> 2) desirable,
> 3) disregarded.
>
> For discussions of association and causation, he cites work on path analysis.

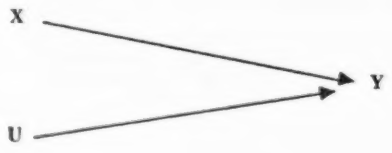
In the next section, I give a brief view of path analysis and suggest environmental statisticians build on the vast experience with this method of analysis.

Environmental Path Analysis– Blazing New Trails with Path Analysis

Why path analysis? Path analysis was developed to solve complex problems in genetics. Path analysis helps in formulation of complex environmental problems and provides a conceptual framework for dealing with complex problems.

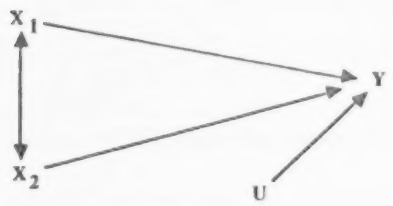
In path analysis one graphically depicts causal pathways to represent statistical relationships among variables.

A simple example might depict a linear relationship between x and y as follows:



Arrow shows direction of the relationship and U represents the unexplained sources of variation in Y.

The path diagram for multiple regression might be drawn as shown below:



Path analysis extends single and multiple regression equation to networks of variables involving more than one equation. Some writers omit U, the unexplained variation, and the correlation arrows between x's from path diagrams.

According to Ching Chun Li (1975) in Path Analysis—a primer, The Boxwood Press, Pacific Grove, CA p. 166

> "When properly analyzed, each path diagram will yield a set of consistent conclusions for that particular causal system."

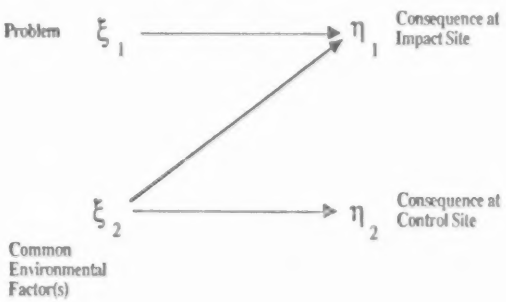
For each environmental problem similar to those described by Alan Stuart-Oaten,

η_1 might denote a site of concern or impact,

η_2 a control site

ξ_1 a problem that affects site of concern, η_1, but not control site η_2

A more realistic model includes common effect of other variables, for example, ξ_2 on both the sites



In conclusion, path analysis provides a way to very quickly build complex models involving many variables. Some authors have extended path analysis to nonlinear models. Nonlinear path models allow investigator a broad range of modeling possibilities and a framework for organizing their work. References are provided to assist the interested reader with the basic concepts of path analysis.

Disclaimer:
The opinions and assertions expressed in this paper are the private conclusions of the author and should not be construed as official or reflecting the views of the U. S. Environmental Protection Agency.

References:

Sewall Wright. Evolution and the Genetics of Populations, Volume 1 Genetic And Biometric Foundations:
Chapter 13. Path Analysis: Theory, pp. 299–324.
Chapter 14. Interpretation by Path Analysis, pp. 325–372.

S. Wright (1921) Correlation and causation. **Jour. Agric. Res. 20**: 557-85.

J. W. Tukey (1954). Causation, regression and path analysis. In Statistics and Mathematics in Biology, eds. O. Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush, pp.35-66. Ames: Iowa State University Press.

M. E. Turner and C. E. Stevens (1959). The regression analysis of causal paths. Biometrics 15: 236–258.

Ching Chun Li (1975) Path Analysis—a primer, The Boxwood Press, Pacific Grove, CA

Studies in Econometric Method by Cowles Commission Research, Edited by Wm. C. Hood and Tjalling C. Koopmans (1953):
Chapter II. Identification Problems in Economic Model Construction.
By Tjalling C. Koopmans, pp. 27-48

Chapter III. Causal Ordering and Identifiability.
By Herbert A. Simon, pp.49–74

APPENDIX A: Program

ASA/EPA CONFERENCE ON SAMPLING AND SITE SELECTION IN ENVIRONMENTAL STUDIES

May 14–15, 1987

Thursday, May 14

8:50 a.m. INTRODUCTION
Walter S. Liggett, National Bureau of Standards
Dorothy Wellington, U.S. Environmental Protection Agency

I. THE STATISTICAL BASIS: RANDOMIZATION AND PROCESS CONTROL

9:00 a.m. Some Statistical Issues in the Collection and Analysis of Environmental Data
George C. Tiao, University of Chicago

10:00 a.m. BREAK

10:10 a.m. Sampling Design: Some Very Practical Considerations
Douglas E. Splitstone, IT Corporation

11:10 a.m. DISCUSSION
Maurice E.B. Owens, Science Applications International Corporation
W. Barnes Johnson, U.S. Environmental Protection Agency

12:10 p.m. LUNCH

II. INFERENCE ON CONTINUOUS SPATIAL DISTRIBUTIONS

1:30 p.m. Spatial Prediction
Noel A.C. Cressie, Iowa State University

2:30 p.m. BREAK

2:40 p.m. Spatial Autocorrelation: Implication for Sampling and Estimation
Evan Englund, U.S. Environmental Protection Agency, EMSL–Las Vegas

3:40 p.m. DISCUSSION
John Warren, U.S. Environmental Protection Agency

4:00 p.m. RECEPTION

Friday, May 15

III. DESIGNS BASED ON AUXILIARY INFORMATION

9:00 a.m. Sampling and Modeling Superfund Site Pollutant Plumes: Methods Combining Direct Measurements and Remote Sensing Data
Ralph E. Folsom, Research Triangle Institute

10:00 a.m. BREAK

10:10 a.m. "Validation" of Air Pollution Dispersion Models
Anthony D. Thrall, Electric Power Research Institute

11:10 a.m. DISCUSSION
Robert W. Jernigan, The American University
Richard A. Livingston, U.S. Environmental Protection Agency and University of Maryland

12:10 noon LUNCH

IV. STATISTICAL COMPARISON OF SITES

1:30 p.m. Assessing Effects on Fluctuating Populations: Tests, Diagnostics, and Decisions
Allan Stewart-Oaten, University of California-Santa Barbara

2:30 p.m. BREAK

2:40 p.m. Comparison with Background Environment: Strategies for Design
Bruce Peterson, CH2M Hill

3:40 p.m. DISCUSSION
R. Clifton Bailey, U.S. Environmental Protection Agency

4:10 p.m. CLOSING REMARKS
Walter S. Liggett, National Bureau of Standards

This Conference is the third in a series of research conferences organized by the American Statistical Association, supported by a cooperative agreement between ASA and the Office of Standards and Regulations, under the Assistant Administrator for Policy Planning and Evaluation, U.S. Environmental Protection Agency.

Conference Chair and Organizer:
Walter S. Liggett, National Bureau of Standards

APPENDIX B: Conference Participants

ASA/EPA CONFERENCE ON SAMPLING AND SITE SELECTION IN ENVIRONMENTAL STUDIES

May 14–15, 1987

Holiday Inn Capital
Washington, D.C.

Joe Abe
U.S. EPA
401 M Street, S.W. (WH–565E)
Washington, DC 20460

Ruth H. Allen
U.S. EPA
Research and Development
401 M Street, S.W.
Washington, DC 20460

Glenn F. Atkinson
Environment Canada
Fifth Floor, Place Vincent Massey
Ottawa, Canada K1A 0H3

R. Clifton Bailey
U.S. EPA
401 M Street, S.W. (WH–586)
Washington, DC 20460

James C. Baker
U.S. EPA, Region 8
999 18th Street, Suite 501
Denver, CO 80202–2405

Ted O. Berner
Battelle Columbus Division
2030 M Street, N.W., Suite 700
Washington, DC 20036

Gordon D. Booth
USDA Forest Service
797 East 5050 South
Ogden, UT 84403

Jill J. Braden
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Dennis L. Brandon
USAE Waterways Experiment Station
P.O. Box 631
Vicksburg, MS 39180

Christine M. Bunck
U.S. Department of Interior
Patuxent Wildlife Research Center
8102 Triple Crown Road
Bowie, MD 20715

Thomas J. Bzik
Air Products and Chemicals, Inc.
P.O. Box 538
Allentown, PA 18105

Jean Chesson
Bertram Price Associates, Inc.
2475 Virginia Avenue, #213
Washington, DC 20037

Robert P. Clickner
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Margaret G. Conomos
U.S. EPA
Exposure Evaluation Div. (TS–798)
401 M Street, S.W.
Washington, DC 20460

Gerald F. Cotton
NOAA/U.S. Department of Commerce
Air Resources Laboratory, R/E/AR
8066 13th Street
Silver Spring, MD 20910

James Craig
U.S. EPA (WH–563)
401 M Street, S.W. Room 2817 WSM
Washington, DC 20460

Noel A.C. Cressie
Iowa State University
Department of Statistics
Ames, IA 50011

James M. Daley
U.S. EPA
Office of Standards & Regulations
Washington, DC 20460

Stephen K. Dietz
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Don Edwards
Department of Statistics
University of South Carolina
Columbia, SC 29208

Evan Englund
U.S. EPA , EMSL–Las Vegas
P.O. Box 15027
Las Vegas, NV 89114

Barrett P. Eynon
SRI International
333 Ravenscross Avenue
Menlo Park, CA 94025

Paul Flyer
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Ralph E. Folsom
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Ruth E. Foster
U.S. EPA
401 M Street, S.W. (PM–223)
Washington, DC 20460

Robert Fusaro
Columbia University
P.O. Box 166, Route 9D
Garrison, NY 10524

Paul H. Geissler
U.S. Department of Interior
Patuxent Wildlife Research Center
12504 Windover Turn
Bowie, MD 20715

Michael Ginevan
Environ Corporation
1000 Potomac Street, N.W.
Washingon, DC 20007

Leigh Harrington
ERIM
1501 Wilson Boulevard, #1105
Arlington, VA 22209

Cynthia L. Hertzler
EG&G Idaho
P.O. Box 1625
Idaho Falls, ID 83415

Todd Higgins
U.S. Army Corps of Engineers
P.O. Box 631
Waterways Experiment Station
Vickburg, MS 39180

W. Ted Hinds
U.S. EPA
401 M Street, S.W. (RD–680)
Washington, DC 20460

Matthew Hnatov
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

Thomas T. Holloway
U.S. EPA
Region 7 Laboratory
25 Funston Road
Kansas City, KS 66115

Robert W. Jernigan
The American University
Math Dept., Clark Hall, Rm. 215
4400 Massachusetts Avenue, N.W.
Washington, DC 20016

W. Barnes Johnson
U.S. EPA
401 M Street, S.W. (PM–223)
Washington, DC 20460

Henry Kahn
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

Louise C. Kern
Texaco, Inc.
P.O. Box 1404
Houston, TX 77251–1404

Debra S. Knopman
U.S. Geological Survey
MS 410, National Center
Reston, VA 22092

Charles T. Kufs
Roy F. Weston, Inc.
5–2 Weston Way
West Chester, PA 19380

Herbert Lacayo
U.S. EPA
401 M Street, S.W., (PM–223)
Washington, DC 20460

Emanuel Landau
American Public Health Assn
1015 15th Street, N.W.
Washington, DC 20005

Barbara Leczynski
Battelle Columbus Division
2030 M Street, N.W., Suite 700
Washington, DC 20036

Walter S. Liggett
National Bureau of Standards
Administration Building, A337
Gaithersburg, MD 20899

Richard A. Livingston
University of Maryland
Geology Department
College Park, MD 20742

Gordon R. Luster
Woodward–Clyde Consultants
100 Pringle Avenue, Suite 300
Walnut Creek, CA 94596–3564

David Marker
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Raymond E. McIntyre
Fairfax County Air Pollution
Control
10777 Main Street, Suite 100A
Fairfax, VA 22030

Daniel I. Michael
Research Triangle Institute
1717 Massachusetts Avenue, N.W.
Suite 102
Washington, DC 20036

John Michael
Westat, Inc.
11725 Happy Choice Lane
Gaithersburg, MD 20878

Steven P. Millard
University of Washington
Biostat. Consulting Unit, SB–77
Seattle, WA 98105

Patricia A. Mundy
U.S. EPA
RD–680, 401 M Street S.W.
Washington, DC 20460

Patricia Murray
U.S. EPA (WH–563)
401 M Street, S.W., Room SE264 WSM
Washington, DC 20460

Avis D. Newell
Northrop Services
Environmental Research Lab
200 SW 35th Street
Corvallis, OR 97333

Maurice E.B. Owens
Science Applications Int'l Corp.
8400 West Park Drive
McLean, VA 22102

Tim B. Parkin
U.S. Department of Agriculture
ARS, Building 007, Room 229, BARC–W
Beltsville, MD 20705

Bruce Peterson
CH2M Hill
P.O. Box 91500
Bellevue, WA 98009

Kurt H. Riiters
Forest Response Program
299 SW 35th Street
Corvallis, OR 97333

Douglas S. Robson
Cornell University
Biometrics Unit, Warren Hall
Ithaca, NY 14853

John W. Rogers
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Charles A. Rohde
Johns Hopkins University
Department of Biostatistics
615 N. Wolfe Street
Baltimore, MD 21205–3179

Marie L. Saint–Louis
EG&G Idaho, Inc.
P.O. Box 1625
Idaho Falls, ID 83415

Paul D. Sampson
University of Washington
Department of Statistics, GN–22
Seattle, WA 98195

Paul A. Schiffelbein
DuPont Company
Engineering Dept., Louviers 51N13
P.O. Box 6090
Newark, DE 19714–6090

Bradley D. Schultz
U.S. EPA (TS–798)
Design Development Branch
401 M Street, S.W., Room E309
Washington, DC 20460

Luther A. Smith
North Carolina State University
Atmospheric Impacts Research Program
2501 Anne Carol Court
Raleigh, NC 27603

Carol J. Spease
U.S. Bureau of Labor Statistics
12405 Braxfield Court, #14
Rockville, MD 20852

Douglas E. Splitstone
IT Corporation
10 Duff Road
Pittsburgh, PA 15235

Stephen V. Stehman
Cornell University
337 Warren Hall
Ithaca, NY 14853

Allan Stewart-Oaten
University of California
Department of Biology
Santa Barbara, CA 93106

Therese A. Stukel
Dartmouth Medical School
HB 7927
Hanover, NH 03756

Clayton L. Stunkard
U.S. EPA
Statistical Policy Branch
401 M Street, S.W.
Washington, DC 20460

Charles Taillie
Pennsylvania State University
CSEES
316 Pond Lab
University Park, PA 16802

Cassie Thompson
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Anthony D. Thrall
Electric Power Research Institute
3412 Hillview Avenue
Palo Alto, CA 94303

George C. Tiao
University of Chicago
Graduate School of Business
1101 E. 58th Street
Chicago, IL 60637

Peter Tong
U.S. EPA
401 M Street, S.W. (WH-548A)
Washington, DC 20460

Y.K. Tung
University of Wyoming
P.O. Box 3322 University Station
Laramie, WY 82071

Alta Turner
EBASCO Services
160 Chubb Avenue
Lyndhurst, NJ 07071

Paul Wakim
American Petroleum Institute
1220 L Street, N.W.
Washington, DC 20005

Elbert Walker
New Mexico State University
Mathematics Department
Las Cruces, NM 88003

John Warren
U.S. EPA
401 M Street, S.W. (PM-223)
Washington, DC 20460

Herbert I. Weisberg
Consulting Statisticians, Inc.
20 William Street
Wellesley, MA 02181

Dorothy Wellington
U.S. EPA
401 M Street, S.W. (PM-223)
Washington, DC 20460

Louis Wijnberg
PEI Associates Inc.
310 Blue Ridge Road
Carrboro, NC 27510

Jack W. Zolyniak
Martin Marietta Energy Systems
P.O. Box P, K-1004D, MS-278
Oak Ridge, TN 37831

American Statistical Asociation

Fred C. Leone
Executive Director

Mary Esther Barnes
Conference Coordinator